TRANSCRIPTOME-WIDE DISCOVERY AND DYNAMICS OF METHYL-6-ADENOSINE

A Dissertation

Presented to the Faculty of the Weill Cornell Graduate School

of Medical Sciences

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Yogesh Saletore

May 2015

TRANSCRIPTOME-WIDE DISCOVERY AND DYNAMICS OF METHYL-6-ADENOSINE

Yogesh Saletore, Ph.D.

Cornell University 2015

RNA modifications have been known to exist since the late 1960s, but the methods to detect them globally did not exist until recently. The release of an antibody specific to the RNA modification methyl-6-adenosine (m^6A) enabled its transcriptome-wide mapping sites using MeRIP-Seq, a variation of RIP-seq that enriches for RNA fragments using the antibody in a pulldown assay and identifies them using next-generation sequencing. The dynamic epitranscriptome has the potential to be involved in multiple layers of regulation, from translation repression to nuclear export and splicing. The purpose of this dissertation was to develop methods both experimentally and computationally to map and identify m^6A sites. As a dynamic modification, m^6A and other RNA modifications change in frequency in response to cell stress and stimuli. Developing methods to detect these changes further elucidate its functional role.

First, the limitations of the MeRIP-Seq protocol were demonstrated, from its input requirements to the consequences of batch effects, ribosomal RNA contamination, and IP efficiency. Second, computational methods were developed to analyze MeRIP-seq data, implementing MeRIPPeR as an m^6A peak finder that performs with higher sensitivity than other peak finders. The consequences of choice of aligner and annotation were also discussed, as well as methods to correct for technical variation and batch effects introduced during the MeRIP-seq protocol. Third, additional computational methods were

developed to identify changes in methylation sites, identifying differentially methylated peak regions to unravel the dynamics of m^6A.

These three methods were then applied to two case studies. In the first study, changes in m^6A sites in response to heat shock and ribavirin treatment in the context of nuclear export were examined. The results showed a correlation between differentially methylated peak regions in introns and changes in nuclear/cytosolic ratios. The second study explored m^6A's role in adipogenesis in the porcine model, serving as the first study of m^6A in pigs. Differentially methylated peak regions found in both studies implicate dynamic m^6A sites in genes involved in RNA regulation, splicing, and nuclear export pathways. The results demonstrate the biological importance of m^6A and further implicate it in RNA processing and regulation.

BIOGRAPHICAL SKETCH

Yogesh Saletore was born in Urbana-Champagne, IL. From there he moved to Corvallis, OR and later to Olympia, WA, where he attended Olympia High School. He graduated with from the University of Washington with dual degrees in Bachelors of Science and Honors in Computer Science and Engineering and Bachelors of Science in Bioengineering in 2010. From there, he joined the Tri-Institutional Computational Biology and Medicine Program and spent one year in Ithaca, NY attending Cornell University. Yogesh joined the Christopher Mason Lab in mid-2011 to pursue his doctoral research in transcriptome-wide mapping and unraveling the dynamics of the RNA modification methyl-6-adenosine.

Dedicated to my parents, my brother Kishore, and Mohana Roy

ACKNOWLEDGEMENTS

I would first like to thank my PhD advisor, Dr. Christopher E. Mason, for his continued support and drive for success. His energy and enthusiasm for science and research has always been an inspiration for my own graduate work. He encouraged me to pursue wet lab/bench side research and helped support me as I struggled in my first few experiments. His vision to be a scientist on both the wet and dry (computational) sides of science has helped me gain a better understanding of the science and methods that we employ. I would also like to thank Dr. Olivier Elemento, Dr. Christina Leslie, and Dr. Adam Siepel for their continued support in my graduate work. I was fortunate enough to take two classes from Dr. Siepel, which inspired me to look into more sophisticated statistical methodologies. My rotation in Dr. Elemento's laboratory was one of my first forays into computational biology and he helped me gain my footing and apply new methods. Dr. Leslie's expertise in CLIP-Seq and machine learning has helped me to develop better statistical models and methods in my research.

I am also grateful to the Tri-Institutional Program in Computational Biology, especially Ms. Margie Hinonangan-Mendoza, Ms. Kathleen Pickering, and Dr. David Christini, for their continued support of my graduate studies. When I first applied to the program in 2010, I was considering mostly computer science programs, but the flexibility and support in the program helped me choose a computational biology program. I am also thankful to the Physiology, Biophysics, and Systems Biology (PBSB) program at Weill Cornell Medical College (WCMC) for accepting me as a part of the group, and especially Ms. Audrey J. Rivera and Ms. Elaine Almonte for their support. I would also like to thank Mr. Jason Banfelder, Ms. Vanessa Borcherding, Mr. Peter Gonzalez, and

the rest of the PB Tech staff for helping me in my computational work, and the Epigenomics Core for their help in sequencing the samples.

I would also like to thank the other members of the Mason laboratory, especially Mr. Paul Zumbo, who taught me how to perform bench side research and would challenge my hypotheses and push me to be an independent thinker. Although he left the lab a few years ago, he has continued to support me, especially in troubleshooting bench side experiments that failed. Mr. Dhruva Chandramohan helped me understand the mathematics and algorithms. I would also like to thank Dr. Sheng Li, Dr. Altuna Akalin, Dr. Cem Mayden, Ms. Priyanka Vijay, Mr. Marjan Bozinoski, Ms. Lenore Pipes, Ms. Heather Geiger, Dr. Elizabeth Hénaff, Dr. Virgínia Mara de Deus Wagatsuma, and Mr. Jorge Gandara, Jr.

I would like to thank Dr. Ari Melnick for helping me in refine the MeRIP-Seq protocol and pushing me to lower the input limits. His insight in antibody work and working with clinical cancer samples has pushed me to further pursue m^6A research in the context of cancer. I am especially grateful to Dr. Tharu Fernando from the Melnick laboratory for her help in growing and isolating cells, nearly 450 million for the IP input test and an additional 1.4 billion for the heat shock and ribavirin experiments. I would also like to thank Dr. Leandro Cerchietti for his ideas and support in developing the Ribavirin and heat shock experimental design.

I am grateful for the continued support of my friends, Dr. Mitchell Kim and Mr. Drazen Doutlik, who have helped me get through the more difficult times in graduate school. I am also thankful to the students of the Tri-I CBM program, especially the students from my year, Ms. Sarah Brooks, Mr. Solomon Shenker, Ms. Tanya Nauvel, Mr. B. Arman Aksoy, Ms. Julie Yang, and Dr. Eyal Nitzani.

Last, but not least, I would like to thank my parents, Dr. Vikram Saletore and Ms. Devika Saletore, for always supporting me and encouraging me to pursue research. I am especially grateful to my brother, Mr. Kishore Saletore, who I am very proud of for graduating from high school. I would like to thank Dr. Mohana Roy, who has been my best friend since the day I met her and has always supported me in my graduate studies.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

- **m^6A:** methyl-6-adenosine
- **IP:** immunoprecipitation
- **MeRIP-Seq:** methylated RNA immunoprecipitation sequencing
- **MeRIPPeR:** MeRIP-Seq Peak FindeR
- **^5mC:** 5-methylcytosine
- **^5hmC:** 5-hydroxymethylcytosine
- **RNA:** ribonucleic acid
- **DNA:** deoxyribonucleic acid
- **mRNA:** messenger RNA
- **tRNA:** transfer RNA
- **rRNA:** ribosomal RNA
- **PCR:** polymerase chain reaction
- **cDNA:** complementary DNA
- **FDR:** false discovery rate

CHAPTER 1 INTRODUCTION TO METHYL-6-ADENOSINE

1.1 Background

Although RNA modifications have been known to exist for many years, only recently have new assays and high-throughput sequencing technologies enabled the characterization of these modifications across the transcriptome. This not only opened the door to examining the physiological and functional role of these modifications, but also their importance in the context of diseases and development.

1.1.1 "The Birth of the Epitranscriptome"

Francis Crick theorized the Central Dogma of Molecular Biology in 1956 to show the flow of genetic information, as DNA is transcribed into RNA and the RNA is translated into proteins. This somewhat simple view ignores the multiple layers of regulation and interaction between the three layers, such as chromosomal recombination in DNA, transcription factors and promoters in transcription regulation, and micro-RNAs (miRNAs) in translation suppression. The term epigenetics was first used in 1942 by C. H. Waddington, but its usage has now evolved to refer to the modifications and structure of DNA, a layer "on top" (epi) of the mutations that can occur in the actual DNA bases. This includes the chromatin structure of DNA, histone modifications such as H3K4 mono-methylation and tri-methylation, and DNA modifications like 5-methylcytosine (^{5}mC) and 5-hydroxymethylcytosine (^{5}hmC).

Although the first RNA modifications were discovered in the late 1960s, the inability to map them transcriptome-wide hindered research. (Iwanami and Brown, 1968) The five-prime cap on mRNA, a methylated guanosine, or 7-methylguanylate, was discovered to regulate nuclear export (Lewis and

Izaurralde, 1997; Visa et al., 1996) and prevent degradation by exonucleases (Burkard and Butler, 2000; Evdokimova et al., 2001; Gao et al., 2000). Initial studies of the most prevalent mRNA modification, N-6-methyladenosine, or methyl-6-adenosine (m^6A) (Figure 1.1), utilized ^{14}C-radiolabeled methionine to observe its incorporation into RNA methyl groups through the endogenous methyl donor, S-adenosylmethionine. They discovered m^6A was present in ribosomal RNA (rRNA) (Iwanami and Brown, 1968), small RNA fractions (Bringmann and Luhrmann, 1987; Desrosiers et al., 1974; Epstein et al., 1980; Levis and Penman, 1978; Wei et al., 1976), and mRNAs (Horowitz et al., 1984).

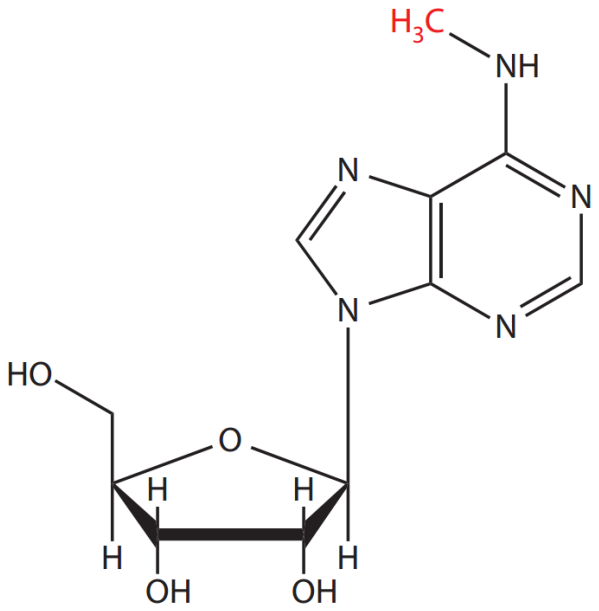


Figure 1.1: Methyl-6-Adenosine (m^6A)
Methyl-6-adenosine with methyl group highlighted in red. Adapted from (Li and Mason, 2014).

Examination of specific methylation sites were restricted to bovine prolactin (Chen-Kiang et al., 1979) and the Rous sarcoma virus (Beemon and

Keith, 1977; Kane and Beemon, 1985). Few groups were interested in m^{6}A until recently when the *fat mass and obesity-associated* (FTO) gene was discovered to also be one of the demethylases of m^{6}A. (Jia et al., 2011) Variants of FTO had previously been linked to obesity (Frayling et al., 2007; Yang et al., 2012) and also Alzheimer's disease (Benedict et al., 2011; Keller et al., 2011), bringing new attention to an almost forgotten field.

An antibody specific to m^{6}A was developed in 1977 by Synaptic Systems but was not made publicly available until recently. (Munns et al., 1977) Following its release, New England Biolabs also made their m^{6}A antibody publicly available (Kong et al., 2000), and now at least five such antibodies have been developed by multiple companies. In collaboration with Kate Meyer, PhD, and Samie Jaffrey, MD PhD, we developed MeRIP-Seq, methylated RNA immunoprecipitation sequencing, to identify m^{6}A sites throughout RNA. Similar to ChIP-Seq (chromatin immunoprecipitation sequencing), the antibody is used to pulldown and enrich for RNA fragments containing m^{6}A, which are then identified using next-generation sequencing. By comparing the enrichment of these fragments to an RNA-Seq control, putative m^{6}A peak regions can be identified. With the establishment of this protocol, m^{6}A could now be studied transcriptome-wide.

This heralded "the birth of the epitranscriptome," (Saletore et al., 2012) a new portmanteau we coined to encompass the new field of RNA modifications. Previously, groups had struggled with non-specific phrases like "RNA methylome" (Dominissini et al., 2012) or "RNA epigenetics," (He, 2010) neither of which we felt fully encapsulated the new field. Although at the time we meant it to refer to all RNA modifications, our focus remained on m^{6}A. However, the

field itself has grown to now include transcriptome-wide 5-methylcytosine mapping in RNA. Utilizing the same bisulfide treatment used in DNA to identify ^{5}mC sites (Meissner et al., 2005), ^{5}mC was successfully mapped in mRNAs and found to be enriched near Argonaute binding regions. (Squires et al., 2012)

In total, the epitranscriptome has the potential to encompass all of the over 100 RNA modifications listed in the RNA Modifications Database, (Agris et al.; Cantara et al., 2011) but only a small fraction can be mapped across the entire transcriptome, at present. However, it should be noted that there is a new focus on pseudouridine, (Jaffrey, 2014) the most prevalent RNA modification and the C-glycoside isomer of uridine. Mostly found in tRNAs, new technologies such as Pseudouridine-Seq (Schwartz et al., 2014) continue to grow the world of the epitranscriptome.

1.1.2 The History of Methyl-6-Adenosine

N^{6}-methyladenosine, also referred to as methyl-6-adenosine or m^{6}A, is the addition of a methyl-group onto the sixth nitrogen of adenosine in RNA. The modification was first discovered in 1968 in rRNAs. (Iwanami and Brown, 1968) Initial studies identified the modification in viruses (Moss et al., 1977), yeast (Bodi et al., 2010) and corn, (Nichols, 1979) with following studies on yeast in 2002 looking at the potential for the dynamic nature of m^{6}A (Clancy et al., 2002). The majority of m^{6}A sites can be found in mRNAs and long non-coding RNAs, but recently they have also been identified in the far smaller micro RNAs (miRNAs). (Berulava et al., 2015)

1.1.3 The Dynamic World of Methyl-6-Adenosine

As m^{6}A and other epitranscriptomic modifications are co-transcriptionally or post-transcriptionally added, they have the potential to be a dynamic level of

regulation, changing in response to cell stimuli and stresses. The methyltransferase of m^6A, or the protein responsible for methylating adenosine to m^6A, was first identified in 1997 by Bokar et al (Bokar et al., 1997) to be METTL3, also known as MT-A70. Subsequently, the methyltransferase complex was discovered to include METTL14 (Liu et al., 2014) and WTAP (Ping et al., 2014). The FTO gene was first identified as a demethylase of m^6A in 2011, (Jia et al., 2011) followed by ALKBH5 in 2013. (Zheng et al., 2013) Knockdown experiments of METTL3 in HeLa cells led to apoptosis, (Dominissini et al., 2012) demonstrating the physiological significance of m^6A. Initial studies in m^6A also identified a sequence specificity for METTL3, as the RRACH consensus, or GGACU motif, where R = guanosine or adenosine and H = adenosine, cytosine, or uracil, (Harper et al., 1990; Wei and Moss, 1977b) and these results were corroborated by transcriptome-wide studies. (Dominissini et al., 2012; Meyer et al., 2012) The methyltransferases have been dubbed the “writers” and the demethylases the “erasers” of m^6A. (Fu et al., 2014)

Some of the potential “readers” of m^6A, or the proteins that bind to the modification, have been identified to be the YTHDF1-3 proteins. (Dominissini et al., 2012; Wang et al., 2014a) In particular, YTHDF2 was found to bind to m^6A sites and targeting them towards P-bodies for degradation. It is theorized that the other readers may have other purposes than RNA decay. It has been long known that m^6A inhibits the activity of the ADAR enzyme to perform A-to-I (inosine) editing (Veliz et al., 2003), though current assays do not allow for specific validation of these sites and changes. While the exact function of m^6A still remains unclear, it has been implicated in splicing and adipogenesis, (Zhao et al., 2014) regulation in embryonic stem cell development, (Wang et al., 2014b) and the circadian rhythm (Fustin et al., 2013). Without fully knowing its

physiological purpose, its functional role can be elucidated by taking advantage of its dynamic nature: examining how it changes in response to cellular stimuli and stresses, between tissue and cell types, and across time.

CHAPTER 2 PROTOCOLS TO IDENTIFY METHYL-6-ADENOSINE

2.1 Prior Publication and Rights to Reprint

Portions of this chapter first appeared in (Saletore et al., 2012), including Figure 2.4 and Figure 2.5. This manuscript is freely available at Genome Biology under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. Full details regarding the Creative Commons License are available at http://creativecommons.org/licenses/by/2.0.

2.2 Introduction

Although an antibody was devised that was shown to preferentially bind to m^6A in the 1970s, it was not publicly available until far more recently. (Munns et al., 1977) Antibodies are often used in Western or immunoblots to show global levels of the epitope. Capitalizing on the wide-spread application of ChIP-Seq, (Johnson et al., 2007) as well as similar methods in ^{5}hmC detection before the advent of TAB-Seq, (Song et al., 2012) the antibody could also be used to selectively enrich for RNA fragments that contain m^6A sites over those that do not. Combining this method with next generation sequencing (NGS), these fragments could be identified and mapped back to the genome, defining peaks, where putative m^6A sites may exist.

2.3 MeRIP-Seq: Methylated RNA Immunoprecipitation Sequencing

Using western blots, Kate Meyer showed that the antibody bound selectively to m^6A, (Meyer et al., 2012) demonstrating that the polyclonal antibody could be used in a pulldown-type assay. The initial method utilized Invitrogen RiboMinus beads to ribo-deplete RNA, to include RNAs that may not have been poly-adenylated, zinc chloride chemical fragmentation of the purified RNA, and

utilized two-rounds of IP to achieve the highest possible enrichment while balancing sample loss.

2.3.1 Ribosomal Contamination

The high input amount of RNA required for MeRIP-Seq will be discussed in more detail in the next section (2.3.2 Input Requirements), but in most RNA-Seq studies, a rather large amount of RNA is required in general because upwards of 90-95% of the total RNA can be comprised of ribosomal RNA (rRNA). Although m^6A is known to be found in some rRNA sites, most of the focus on its function is in messenger RNA (mRNA) and non-coding RNA (ncRNA), so successful removal of rRNA is crucial. Contamination by rRNA leads to significant loss of sequencing depth in genomic regions that are of interest. The two primary methods of rRNA removal utilize either polyA-pulldown or ribo-depletion, the prior utilizing a string of deoxy-thymines to selectively bind to the poly-adenylated (polyA) tails on mRNAs to pull down and enrich for them, and the latter using cDNA complexes for rRNA sequences to bind to them and knock them down. The advantage of utilizing a ribo-depletion method is to enrich for RNA families that may be missed in a polyA pulldown, such as ncRNAs and long non-coding RNAs (lncRNAs) (Li et al., 2014b; SEQC MACQ-III Consortium, 2014; Yamamoto et al., 2014), but recent studies have shown that non-specific cDNA binding with the rRNA removal complexes can lead to sequencing bias (Lahens et al., 2014).

The original MeRIP-Seq protocol requires upwards of 300 micrograms of total RNA as input, but utilizes RiboMinus™ Human/Mouse Transcriptome Isolation Kit (Life Technologies #K1550-01) to perform the ribo-depletion step. Per the manufacturer instructions, the beads can at most purify only 5 micrograms of

RNA in a single knockdown and cannot be re-used. Consequently, quality control measurements (shown in Figure 2.1) confirm the problem that utilizing such a high total RNA input with the RiboMinus beads results in not only a high degree of rRNA contamination in the sequencing sample, but also high variation of contamination between samples. While the sequencing loss in both money and sequencing depth is enormous, the further consequences of it on peak calling are discussed later in 3.4.5 Ribosomal RNA Contamination. This does demonstrate that for the standard high input protocol, RiboMinus and similar ribosomal knockdown methods are not suitable for full rRNA removal.

One of the advantages of using polyA-pulldown enrichment is their high efficiency in rRNA removal, though at the loss of other RNA species, such as non-coding RNAs and long non-coding RNAs that are not polyA-tailed. However, it should be noted that fully intact and high quality RNA is also required, as mRNAs are pulled down by their polyA tails at their 3' untranslated regions (3' UTRs). Using degraded or low-quality RNA in a polyA-pulldown results in a 3' UTR bias in the data and a loss of 5' UTR coverage. (Li et al., 2014b) That being said, Dynabeads® Oligo$(dT)_{25}$ (Life Technologies #61005) can not only be used to purify 75 micrograms of total RNA per reaction, but they can also be washed and re-used on the same sample again to purify up to a total of 300 micrograms of total RNA. Using these instead of the RiboMinus beads shows a dramatic improvement in rRNA removal, as shown in Figure 2.2. A parallel study in m^6A mapping also successfully utilized the polyA-beads to achieve mRNA purification, (Dominissini et al., 2012) though they unfortunately, and incorrectly, deemed the step as "optional" in their published methods protocol. (Dominissini et al., 2013)

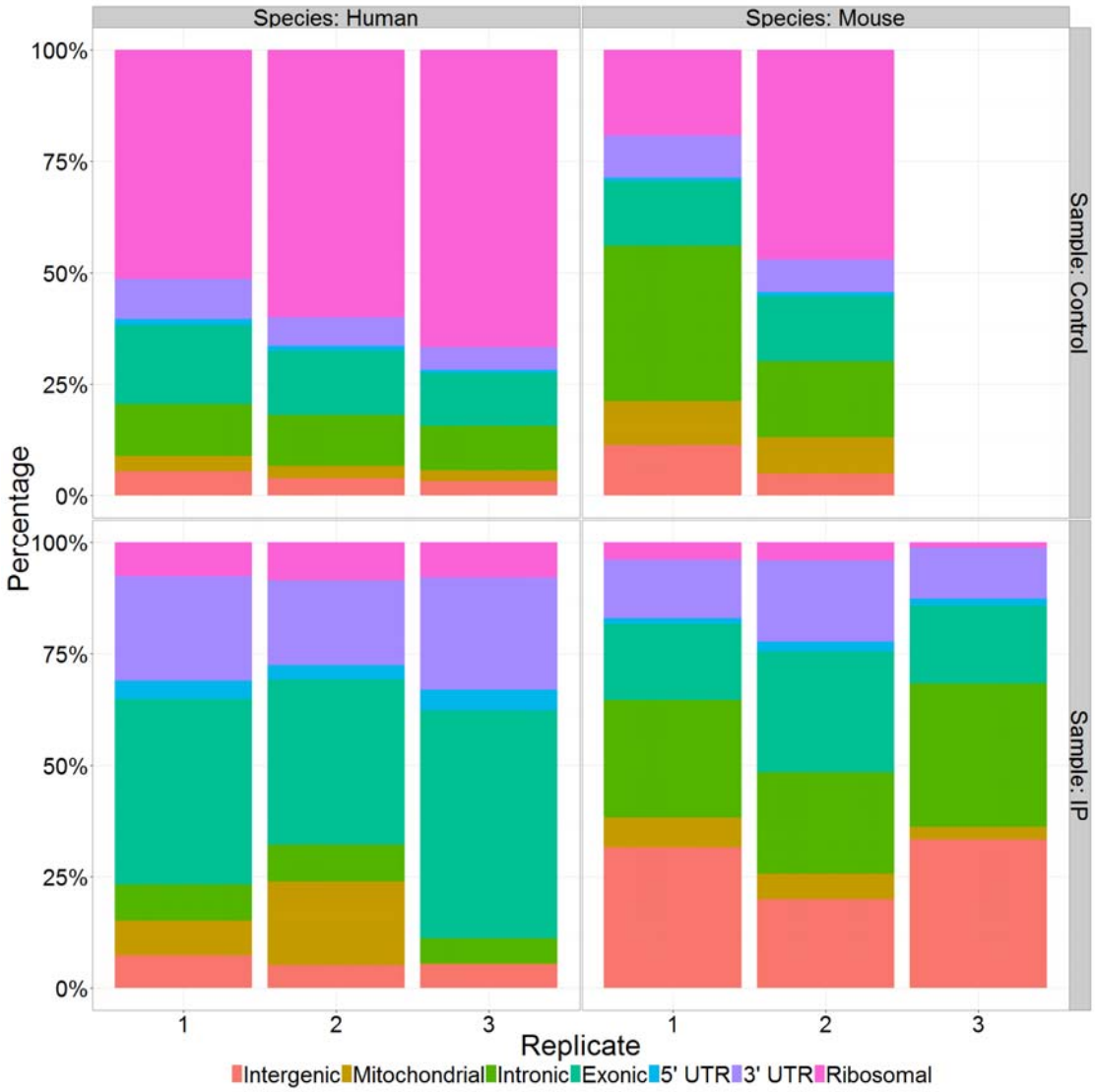


Figure 2.1: Poor Ribo-depletion in Meyer et al. (2012) Samples
The distribution of RNA-sequencing data from Meyer et al (2012) samples, with percentage of reads mapping to intergenic regions in salmon, mitochondrial in dark yellow, intronic in green, exonic in teal, 5' UTR in cyan, 3' UTR in purple, and ribosomal in pink. The human and mouse samples from show a high degree of rRNA contamination, especially in the control samples, and a high variation of contamination between replicates. Reads that would normally be equally distributed across other gene features, such as the exons, are now consumed heavily by ribosomal regions.

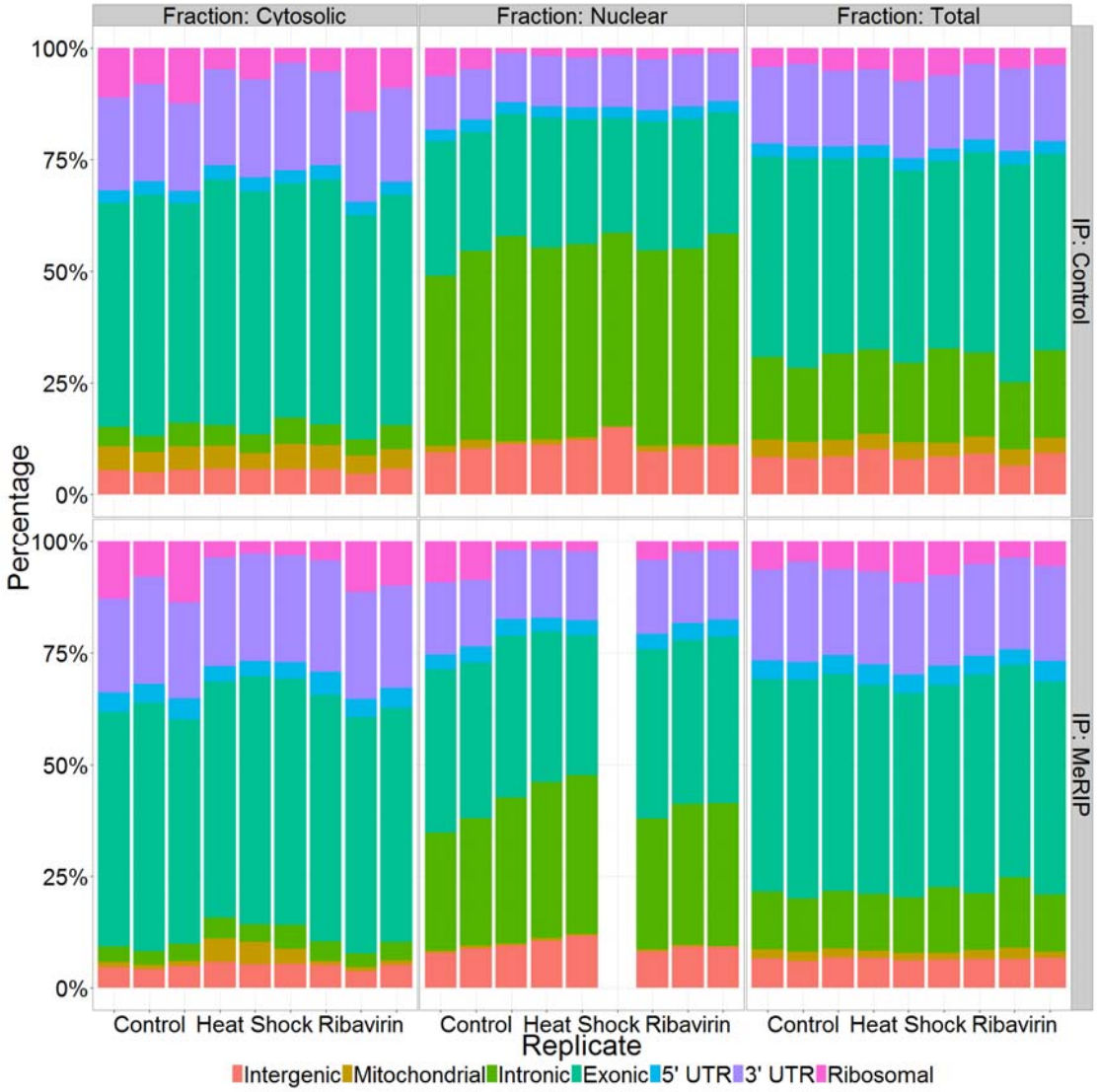


Figure 2.2: RNA-Sequencing Distribution in Samples Utilizing PolyA Pulldown. The distribution of RNA-sequencing data from Heat Shock and Ribavirin samples (Chapter 5), with percentage of reads mapping to intergenic regions in salmon, mitochondrial in dark yellow, intronic in green, exonic in teal, 5' UTR in cyan, 3' UTR in purple, and ribosomal in pink.PolyA-pulldown RNA achieves a significant improvement in rRNA removal over RiboMinus with a high input amount of RNA.

2.3.2 Input Requirements

One of the main challenges to MeRIP-Seq is its high input requirement, in the hundreds of micrograms of total RNA. Methyl-6-adenosine was initially

estimated to be present on 0.1-0.4% of all adenosines, (Dubin and Taylor, 1975; Perry and Scherrer, 1975; Wei et al., 1975) so the amount of RNA being pulled down is very small in even a very successful IP, compounded by the removal of 90% of the total RNA in the rRNA removal step, as discussed previously. This puts MeRIP-Seq out of the range of most clinical samples and requires utilizing either whole organs or cell lines. The original protocol was also designed for two rounds of IP. Using qPCR, one round was estimated to achieve a 75-fold enrichment, and two-rounds a 130-fold enrichment, of m6A fragments over the background. (Meyer et al., 2012) After the publication of the initial paper, the actual limitations of the protocol and its viability for clinical samples needed to be assessed. A titration experiment was performed by varying the input amount of RNA and attempting to perform both one and two rounds of IP until a point at which successful RNA-Seq libraries could not be built from the immunoprecipitated RNA.

Table 2.1 shows the input amounts that were tested, though the fraction of fragmented polyA-purified RNA was estimated to be 1% of the total RNA input. More than 450 million Ly1 cells were collected by Tharu Fernando, PhD from the Ari Melnick, MD laboratory for the purposes of this test. The RNA was extracted, polyA-purified, and then fragmented to generate a consistent pool to not only measure what IP titrations were successful, but also its downstream consequences on sequencing and peak calling. The full titration experimental design would ultimately compare the impact of rounds of IP, input amount, and the library preparation method (Clontech vs Illumina). Table 2.1 also shows which IPs were successful and prepared into libraries with the Clontech and Illumina kits for sequencing.

Table 2.1: MeRIP-Seq RNA Input Titrations and Successful IPs

RNA Input	mRNA	Antibody	1x IP	2x IP
300 µg	3 µg	12 µg	X	X
100 µg	1 µg	12 µg	X	X
50 µg	500 ng	12 µg	X	X
1 µg	10 ng	12 µg	X	X
500 ng	5 ng	12 µg	X	
250 ng	2.5 ng	12 µg	X	
100 ng	1 ng	12 µg	X	

The full experimental design prepared all samples using both Illumina and Clontech kits in two replicates each. In order to eliminate batch effects between preparation methods, samples from a single replicate were immunoprecipitated and pooled, before separating them again into their respective methods. For example, for a single replicate across the design, 4 total IPs were performed with the 100 µg input, pooled, and then separated into four aliquots, two for 1-round IP preparation using Illumina and Clontech protocols, and the other two were immunoprecipitated for a second round, pooled, and then prepped using Illumina and Clontech. Figure 2.3 demonstrates how the RNA samples were separated, fragmented, immunoprecipitated, pooled, and prepped.

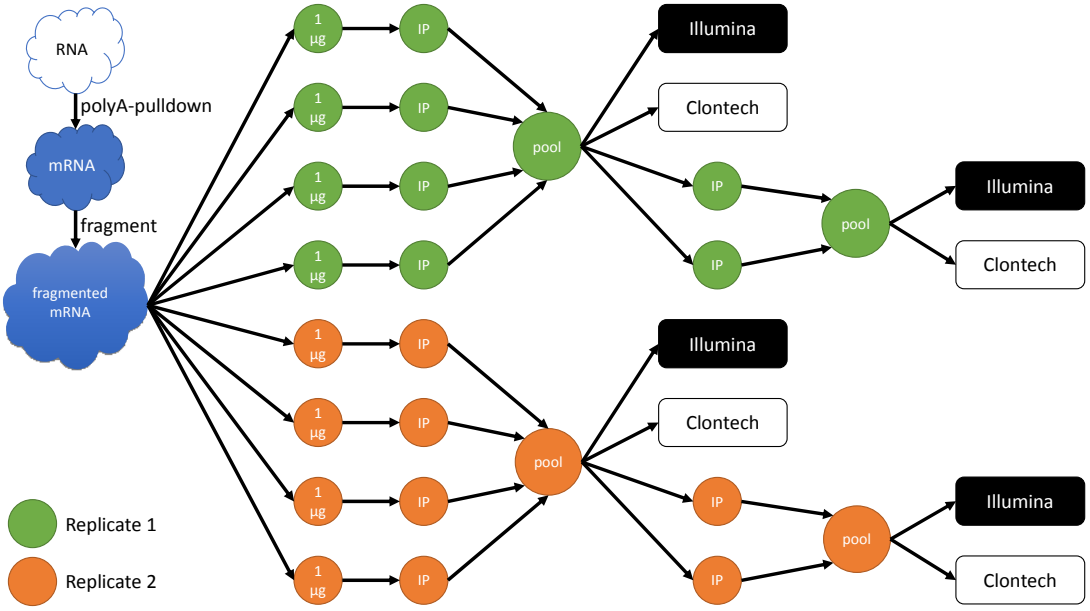


Figure 2.3: 100 µg Input Experimental Design to Minimize Batch Effects
The experimental design for the 100 µg RNA input level shows how samples were extracted, pooled and split to minimize batch effects. The first pooling is to minimize batch effects between the IPs as well as to have a consistent RNA-pool to input into the sequencing prep, and the second pooling is to again maintain the same pool of RNA inputted into Illumina and Clontech preparation methods. Replicates are denoted by color, green and orange, while Illumina prepped-samples are shown in white text on black and Clontech in black text on white background.

Surprisingly, the 2-round IPs succeeded until about 50 µg input, and the 1-round IPs until 100 ng. However, later experiments revealed that this is likely because first, estimating the fragmented mRNA input as one percent of the total RNA was likely an overestimate and does not reflect the amount of RNA lost in each of the steps. Second, the experiment viability also depends highly on the efficiency of the IP achieved, which itself is a function of the antibody purity and function. As the antibody degrades, or if poorer antibodies are used, we continued to observe a drop in the success of IPs that had previously succeeded.

Although the Clontech kit did not provide as much of an advantage over the standard Illumina TruSeq kit, the experiment did provide insight into future experiments with the kit. Upon discussing the results with Clontech technical support, the first lesson was that the Clontech first-strand synthesis reverse transcription enzyme is hindered by glycogen, which unfortunately is used as a carrier in all of the ethanol precipitations. A potential solution is to perform the cleanups blind, without a carrier, but the possibility of sample loss is very high. The second challenge was that Clontech's single-cell range kits are designed to utilize polyA-primed reverse transcription enzymes, which only work with fully intact total RNA. The input to the IP is fragmented RNA, and so random hex priming must be used, which unfortunately does not perform as well as polyA-priming, per Clontech's advice. The kit that was ultimately used was the SMARTer Universal Low Input RNA Kit (Clontech Catalog #634946), originally designed for degraded or formalin-fixed paraffin-embedded (FFPE) samples, which functions by using a random hex primer to generate cDNA complexes, and then utilizing additional PCR cycles, up to a total of potentially 30 cycles, to further amplify the library.

Third, the Clontech manual recommended trimming up to seven nucleotides *in silico* post-sequencing, because some parts of the adapter sequences may remain, even after RsaI digestion. What was not mentioned was that these sequences may in fact impact sequencing color balancing, especially during the first five cycles, when the sequencers require full color balancing to accurately calibrate the bases. This resulted in a low number of reads passing the pass filter cutoff, and a very low sequencing throughput. A possible solution, developed for a similar problem observed in eRRBS (enhanced reduced bisulfide sequencing), is to perform *dark cycle* sequencing, where the bases are

sequenced as normal, but the camera is turned off. The camera can then be turned back on after the first seven base pairs have been skipped to finish the sequencing run. These subsequent bases would then have the necessary complexity for the sequencer to perform accurate base calibration.

2.4 Direct RNA Sequencing

The challenge with mapping m^6A sites at single base-resolution is that current sequencing technologies, referred to as *next-generation sequencing* (NGS) or second-generation sequencing, are dependent on developing cDNA, or complementary DNA, libraries that are ultimately sequenced, after being polymerase chain reaction (PCR) amplified. RRBS (Meissner et al., 2005) takes advantage of bisulfide treatment to convert all non-methylated cytosines to uracil, which is sequenced as a thymine. Since methods to chemically convert m^6A and many other RNA modifications remain unknown, the creation of the cDNA library ultimately results in the loss of all base modification information. One potential solution to achieve single nucleotide resolution is to directly sequence the RNA molecules. Third-generation sequencers was a term first used to generalize sequencing technology that directly sequences the DNA molecules directly, without utilizing any PCR amplification. Traditionally, these methods were designed for direct sequencing of DNA molecules, and RNA sequencing still relies heavily on cDNA library construction. However, research has shown that there still remains the potential to directly sequence RNA transcripts to identify base modifications.

2.4.1 Pacific Biosciences RS

In Pacific Biosciences of California, Inc.'s single-molecule real time (SMRT) technology, DNA polymerases are bound to the bottom of thousands of very

small wells, also known as zero-mode waveguides (ZMWs). A DNA strand is then fed to the polymerase and a camera records the replication of the strand in real time, as fluorescently tagged nucleotides are incorporated into the new strand. The fluorescent pulses can then be converted into base information, directly sequencing the original DNA strands. (Eid et al., 2009) Although the technology has been successful in achieving far longer reads than current Illumina and Ion Torrent technologies, the high error rate of up to 15% of each base has been a challenge and often requires high coverage to achieve a strong consensus sequence.

In observing the incorporations of the nucleotides, PacBio further observed a consistent delay in the incorporation of some of the bases. They realized the inter-pulse distance (IPD), or the amount of time between each fluorescent observation, was dependent on the base modifications present on the original DNA strand. (Flusberg et al., 2010; Song et al., 2011) The activity of the DNA polymerase and its incorporation of nucleotides into the new strand is dependent on both the genetic and epigenetic markers. Although it requires even greater coverage than standard DNA-sequencing to call DNA modifications with high statistical confidence, the method first demonstrated that the technology could be used to discern unmodified bases from modified ones.

The PacBio RNA-sequencing platform currently leverages the Iso-seq protocol, which creates long cDNA templates for sequencing. However, replacing the original DNA polymerase in the ZMW with a Human Immunodeficiency Virus (HIV) reverse transcriptase, native RNA transcripts can instead be fed and directly translated. Instead of performing cDNA synthesis prior to loading the templates onto the machine, the process of generating the cDNA template can

be observed. Comparing two synthetically prepared RNA oligonucleotides, one made with adenosines and the other with m^6A, a similar kinetic signature can be observed as in DNA, as depicted in Figure 2.4 and the distribution shown in Figure 2.5. The study demonstrates the first direct-RNA sequencing method to identify RNA modifications at single nucleotide resolution. However, the figure also shows limitations of the method. The reverse transcriptase “stutters” during cDNA construction, which can be observed as multiple adenosine spikes for the sequencing of a single thymine base. This problem was in fact first observed in DNA sequencing on the PacBio platform, which was solved through adjustments to the DNA polymerase. Similar adjustments to the reverse transcriptase could yield more successful results.

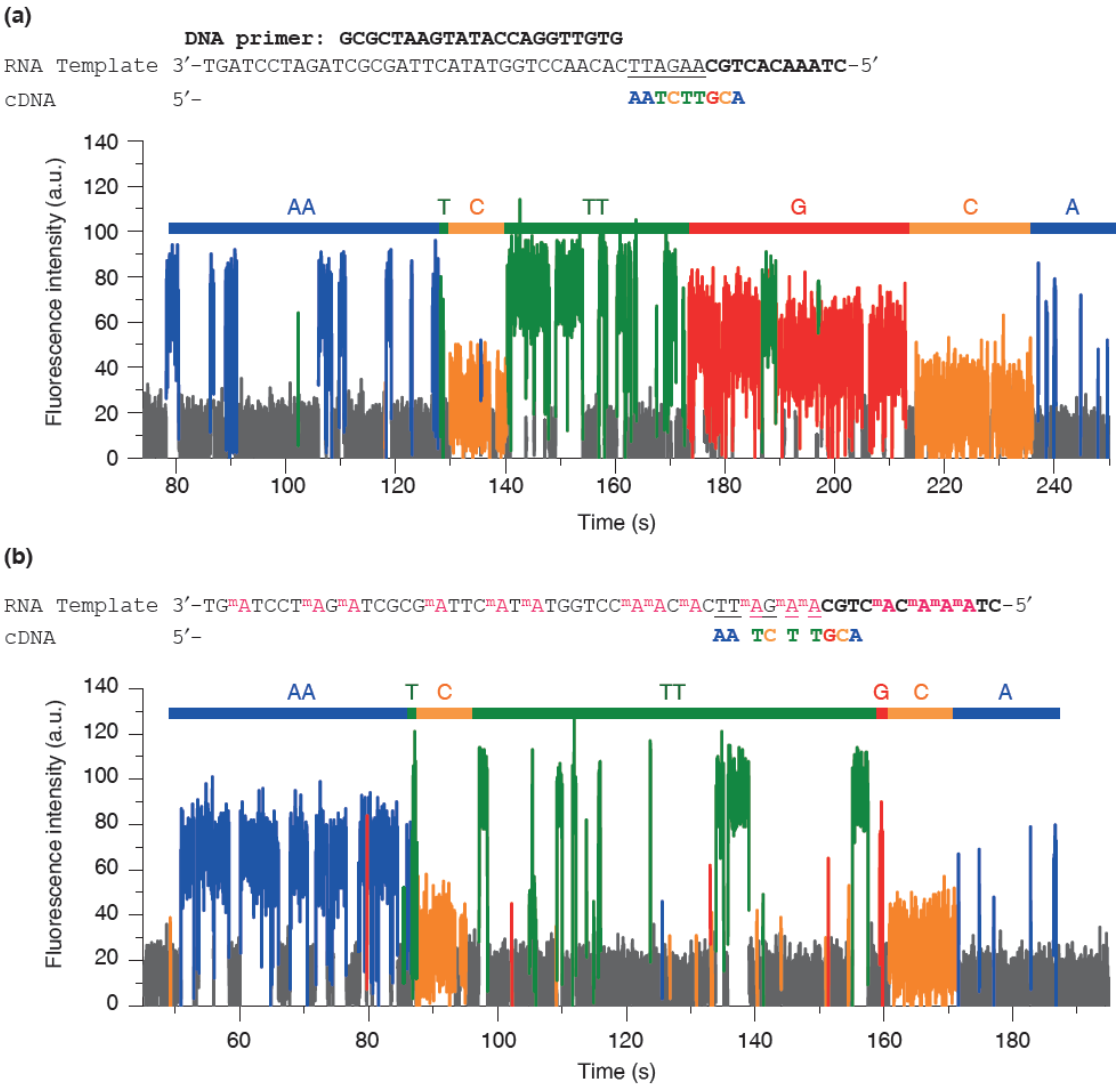


Figure 2.4: Delay in Fluorescent Spikes in Presence of m^6A
Sequencing of synthetic RNA oligonucleotides on the PacBio RS, showing the time of base incorporation and fluorescence intensity in a.u. (normalized arbitrary units) as cDNA is generated by a reverse transcriptase. **(a)** Shows cDNA template construction of the normal oligonucleotide. **(b)** cDNA construction of the modified template, with m^6As substituted for all adenosines in the template used in (a), showing a dramatic delay in the incorporation of thymines corresponding to m^6A sites. (Saletore et al., 2012)

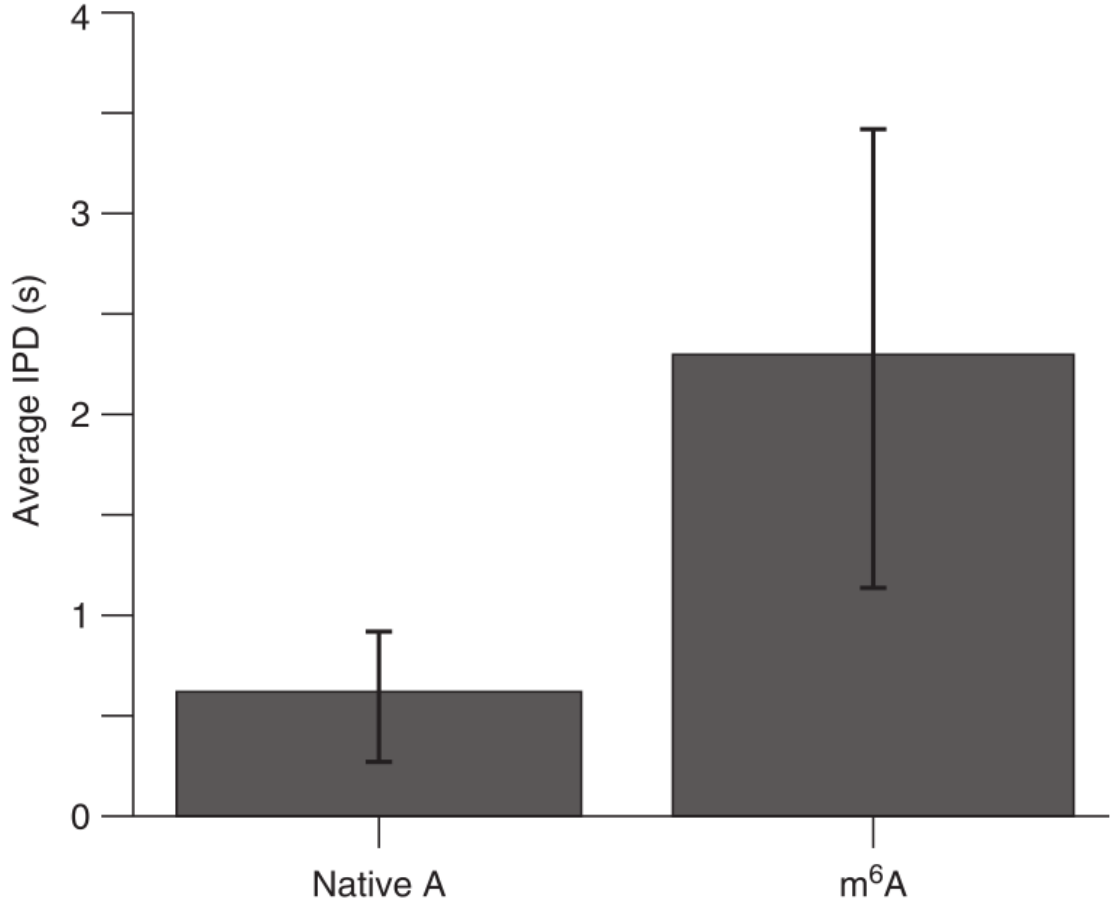


Figure 2.5: m^6A Sites Marked by Increased Inter-Pulse Distance (IPD)
Box plot comparing the average inter-pulse distances for native adenosine (left) and m^6A (right) shows a significant increase in the IPD in m^6A. (Saletore et al., 2012)

2.4.2 Oxford Nanopore Technologies

Most current sequencing technologies take advantage of some form of replication with fluorescently tagged nucleotides, which are then used to reconstruct the original template. In contrast, instead of recording replication, sequencing methods to detect the DNA sequences using nanopores have been theorized and researched since the mid-1990s. (Kasianowicz et al., 1996) A DNA strand is guided through a nanopore with an electrical current passing through it, and changes in the current can be observed as different nucleotides

pass through the pore. Initial experiments showed that these current changes could be measured, but the resolution was low because the DNA moved through the nanopore too quickly to achieve single-nucleotide resolution. (Bayley, 2006)

Oxford Nanopore Technologies licensed the usage of nanopore research from Harvard, UCSC to eventually develop the first commercialized nanopore sequencer. Theorized to use a custom alpha-hemolysin pore, changes in the electrical current as DNA molecules pass through the pore are recorded. Changes in current are recorded, which represent five-base k-mers of the template strand. Using a Hidden Markov-Model (HMM), these k-mers are then converted into bases. Although the average error in a single base can be as high as 30%, and its error model of increased insertions and deletions does not allow for successful alignment with most conventional aligners, the method does show promise in the ability to directly sequence the native DNA strand. DNA fragments are ligated with hairpin adapters that when pulled into the nanopore, can be used to pull the second strand through the pore. This method of sequencing, noted as 2D-sequencing, can be used to achieve higher base quality accuracy.

Oxford Nanopore Technologies' current focus has been on developing their DNA sequencing technology and improving base quality and throughput. However, some of their initial work has indicated that it may be possible to detect DNA modifications at single-nucleotide resolution using the same technology. As the DNA molecules pass through the pore, modified bases do appear to have their own electric signatures. Their current RNA sequencing protocol employs a cDNA library construction step, but previous iterations utilized only first-strand synthesis to generate a cDNA:RNA hybrid molecule that was then sequenced.

The hairpin adapters can be used to pull the RNA strand through, resulting in direct RNA sequencing, similar to what was achieved earlier on the PacBio platform. In particular, this has to potential to detect RNA modifications at single-nucleotide resolution.

2.5 Conclusions and Future Work

Until a chemically-based method, such as using bisulfide treatment to find ^{5}mC, can be formulated to differentiate m^{6}A sites from adenosine, MeRIP-Seq serves as the best method for full transcriptome-wide mapping of m^{6}A sites. The method is dependent on the successful pulldown of RNA fragments using an antibody, introducing high variance and variability into the protocol. The original high input limit was required to ensure successful two-rounds of IP, but for experimental designs with limited RNA input, a single-round can be used to enable IPs at lower inputs. For ultra-low input experiments, below the one microgram range, a method similar to the iChIP protocol could be adapted. (Seumois et al., 2014) Instead of ligating DNA adapters, RNA adapters could be ligated to the RNA fragments, enabling pooling of samples. (Shishkin et al., 2015) While a single sample may not have enough RNA for a successful IP, pooling multiple samples could enable IPs at even lower ranges. Carrier RNA from a cell line could be further used to ensure that the IP is successful.

CHAPTER 3 MERIPPER: MERIP-SEQ PEAK FINDER

3.1 Prior Publication and Rights to Reprint

3.2 Introduction

The MeRIP-Seq protocol comprises the wet laboratory/bench half of identifying m^6A sites, pulling down RNA fragments with m^6A sites and sequencing them. As is the case in ChIP-seq, the computational half is aligning these fragments and converting them into putative m^6A site locations, or peaks. These peaks can then be annotated with known gene locations and used to elucidate the possible functional role of m^6A.

3.2.1 Previous Peak Calling Methods

Since the advent of ChIP-seq protocols, multiple peak finders have been created that use a multitude of methods and statistical tests to attempt to identify peaks. Many of these were specifically designed for ChIP-seq analysis but may still be used to find IP-rich regions in RNA-seq data. Some of them can take as input a control sample, which in theory could be used to some extent to normalize by RNA transcript levels. MACS (Zhang et al., 2008) has become the most commonly-used peak finder in ChIP-seq analysis and was also used to identify peaks in multiple m^6A studies, (Dominissini et al., 2013; Dominissini et

al., 2012) while ChIPseeqer (Giannopoulou and Elemento, 2011) provides a full suite of annotation and motif finding tools in addition to its peak finder.

MeRIPPeR (Meyer et al., 2012; Saletore et al., 2012) was the first m^6A-specific peak finder, designed to use heuristics specific to MeRIP-Seq data. Although its methods were made public, as a stand-alone tool it had not been published and another group adapted a Perl implementation that used the same methods. (Li et al., 2013) Following this, the peak finder exomePeak, (Meng et al., 2013; Meng et al., 2014) which analyzes peaks only within exonic regions, was published as a "FRIP-Seq" (Meng et al., 2013) peak analysis tool, to analyze all fragmented RNA IP sequence data. The software is packaged as a Bioconductor (Gentleman et al., 2004) package, processes windows spanning across connected exons from an inputted annotation, and uses the Poisson distribution to model the read count data.

3.3 Methods

3.3.1 Alignment

The first step in the analysis of any sequencing data is the accurate alignment of the short-reads to a reference genome. In Meyer et al. (2012) the burrows-wheeler genomic DNA aligner BWA (Li and Durbin, 2009) was used and the Dominissini et al study (Dominissini et al., 2012) used the equivalent BowTie aligner (Langmead et al., 2009). At the time, gap and splicing-aware aligners, such as TopHat (Trapnell et al., 2009) and GSNAP (Wu and Nacu, 2010), were still in their infancy, and STAR (Dobin et al., 2013) had not been published. The choice of aligner can have a significant impact on the number of peaks called and the accurate annotation of those peaks. (Saletore et al., 2012) Figure 3.1 shows the fraction of reads successfully mapped and Figure 3.2 shows the

distribution of the mapped reads to gene features, both comparing the choice of aligner and the usage of a gene annotation.

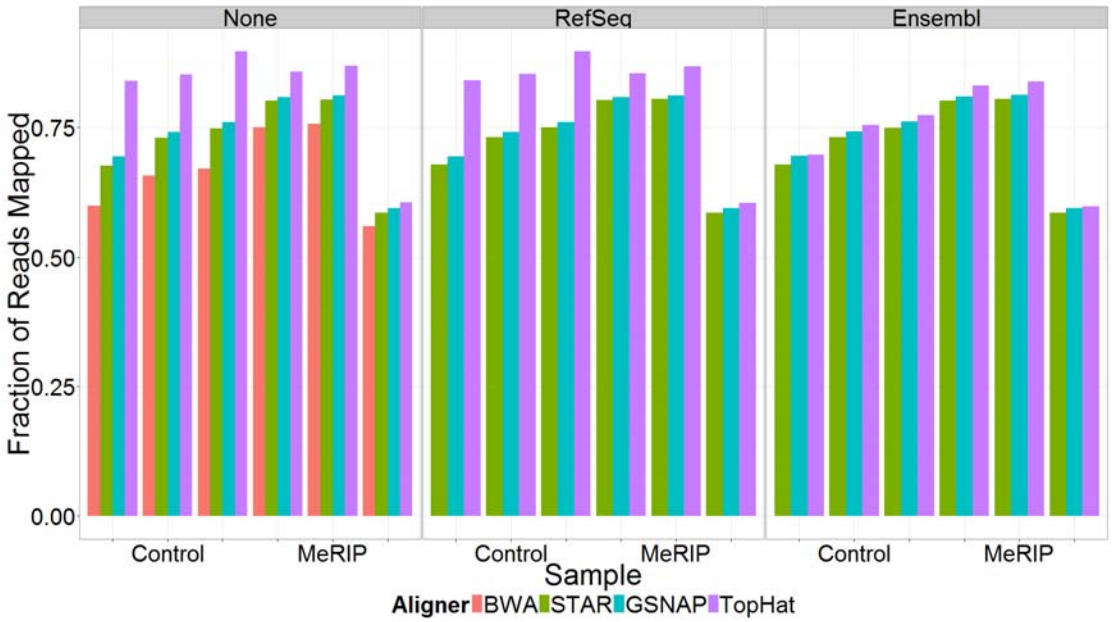


Figure 3.1: Gapped RNA-Seq Aligners Map More Reads than BWA
Fraction of the reads mapped is shown for each of the Meyer et al. (2012) human samples, with BWA shown in peach, STAR in lime-green, GSNAP in cyan, and TopHat in purple. BWA aligns the fewest reads, likely due to its inability to map spliced reads. TopHat aligns more reads compared to the other aligners without an annotation set or using RefSeq annotation, so these aligned reads may potentially represent falsely spliced reads, as fewer reads are mapped using TopHat with Ensembl annotations.

BWA, BowTie, and other genome aligners were specifically designed to align DNA-sequencing data to the reference genome. RNA transcripts in Eukaryotic organisms are assembled from splicing together exons from immature pre-mRNA transcripts and removing intronic segments. (Will and Luhrmann, 2011) This process occurs in spliceosomes in nuclear speckles found in the nucleus, (Lamond and Spector, 2003) incidentally where FTO was found to co-localize, implicating m^6A in this process. (Jia et al., 2011) Splicing complicates

the alignment of reads to a reference genome, specifically reads coming from RNA fragments that span across a splice junction.

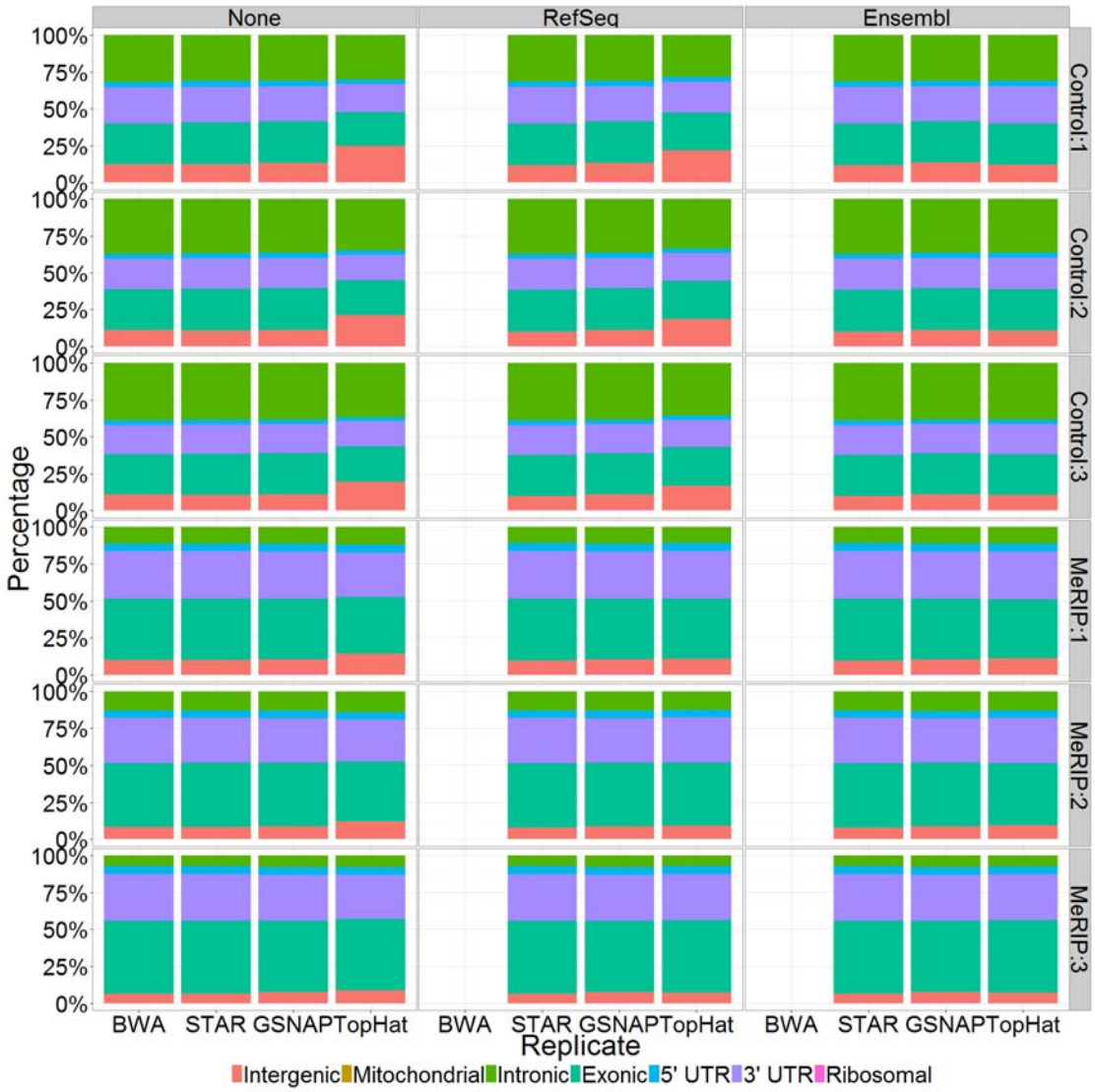


Figure 3.2: TopHat Aligns More Reads to Intergenic Regions without Annotation The distribution of reads mapped by different aligners (x-axis) to gene features by percent (y-axis), with intergenic in salmon, mitochondrial in dark yellow, intronic in green, exonic in teal, 5' UTR in cyan, 3' UTR in purple, and ribosomal in pink. The distributions look mostly comparable between different aligners. TopHat tends to align more reads to intergenic regions without a reference annotation, which could be caused by mis-aligned reads.

Genomic aligners were not designed to handle spliced data, resulting in lack of coverage near splice-junctions, as shown in Figure 3.3, corroborating the fewer reads mapped by BWA in Figure 3.1. DNA-aligners can be used to align RNA-seq data to a reference transcriptome, fixing analysis to an annotation.

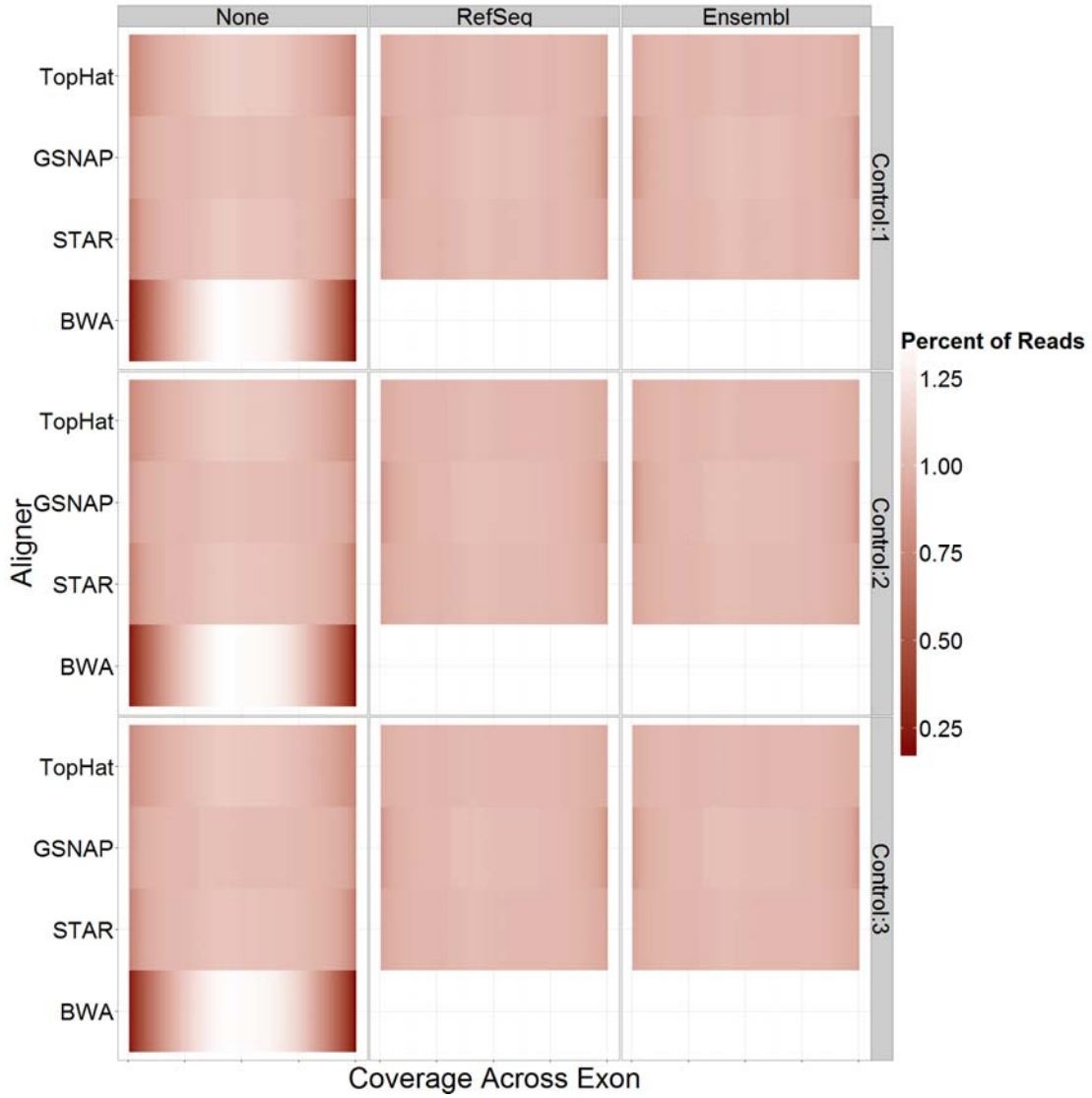


Figure 3.3: BWA Shows Lack of Coverage at Exon Ends
Heat map of percentage of reads mapping to each exon binned into 100 bins on the x-axis, with aligner shown on y-axis, and choice of annotation varying horizontally for the three control samples from Meyer et al. (2012).The splice-aware aligners map reads more uniformly across exons, while BWA has clear drops at the 5' and 3' edges of exons.

Gapped aligners, such as TopHat, GSNAP, and STAR, are specifically designed to map RNA-sequencing data to a reference genome. TopHat works by attempting to align reads using BowTie and then handling those reads that do not map well separately. (Trapnell et al., 2009) STAR aligns maximal mappable prefixes during the seeding phase, and then extending these prefixes to detect splice junctions and mismatches. (Dobin et al., 2013) GSNAP uses k-mers, such as oligomers of 8-mers, to perform a similar alignment. (Wu and Nacu, 2010). STAR is often favored over GSNAP and TopHat, for both its speed and for TopHat's high false positive rate (Figure 3.1 and Figure 3.2). (Li et al., 2014a; Li et al., 2014b; SEQC MACQ-III Consortium, 2014) Accurate alignment of the reads directly affects the ability to call peaks, especially peaks near splice junctions. (Saletore et al., 2012)

Venn diagrams showing the impact of both choice of aligner, Figure 3.4, as well as the impact of using an annotation database on STAR, GSNAP, and TopHat in Figure 3.5, Figure 3.6, and Figure 3.7, respectively. STAR performs the best, with a good balance of speed and accurate alignment, while TopHat has artifacts present both in the distribution of mapped reads and peaks called in the absence of an annotation database. GSNAP has the best agreement with or without using an annotation database, and calls mostly the same peaks as STAR.

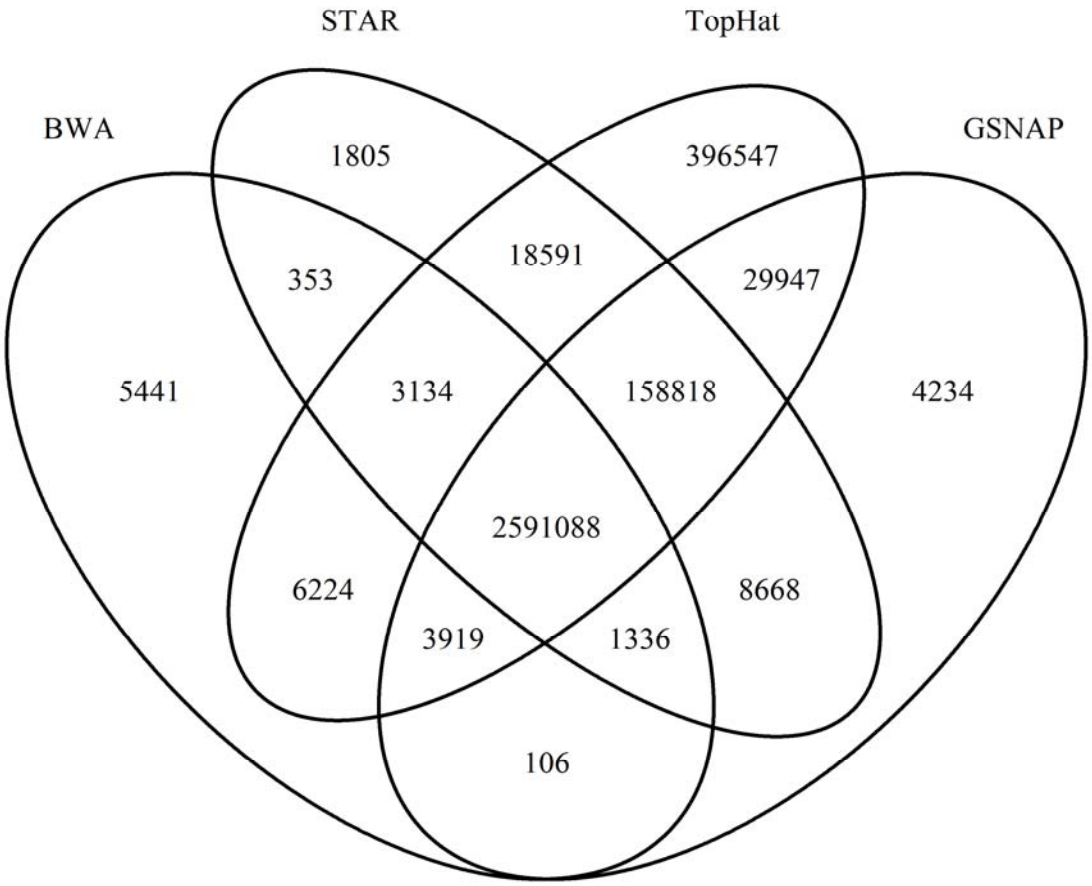


Figure 3.4: Most Bases Common to all Peak Callers
The number of bases called by each of the peak callers in their respective peaks and bases common to overlapping peaks is shown. The different aligners call essentially the same peak regions, with the majority of bases in the intersection of all of them. Depicted is count of the number of base pairs common to each region. BWA misses peaks near exon ends, while STAR has the best balance of speed and accuracy. TopHat has an unusually high number of peaks unique to itself, likely the result of incorrect read alignments. (Li et al., 2014a; Li et al., 2014b; SEQC MACQ-III Consortium, 2014)

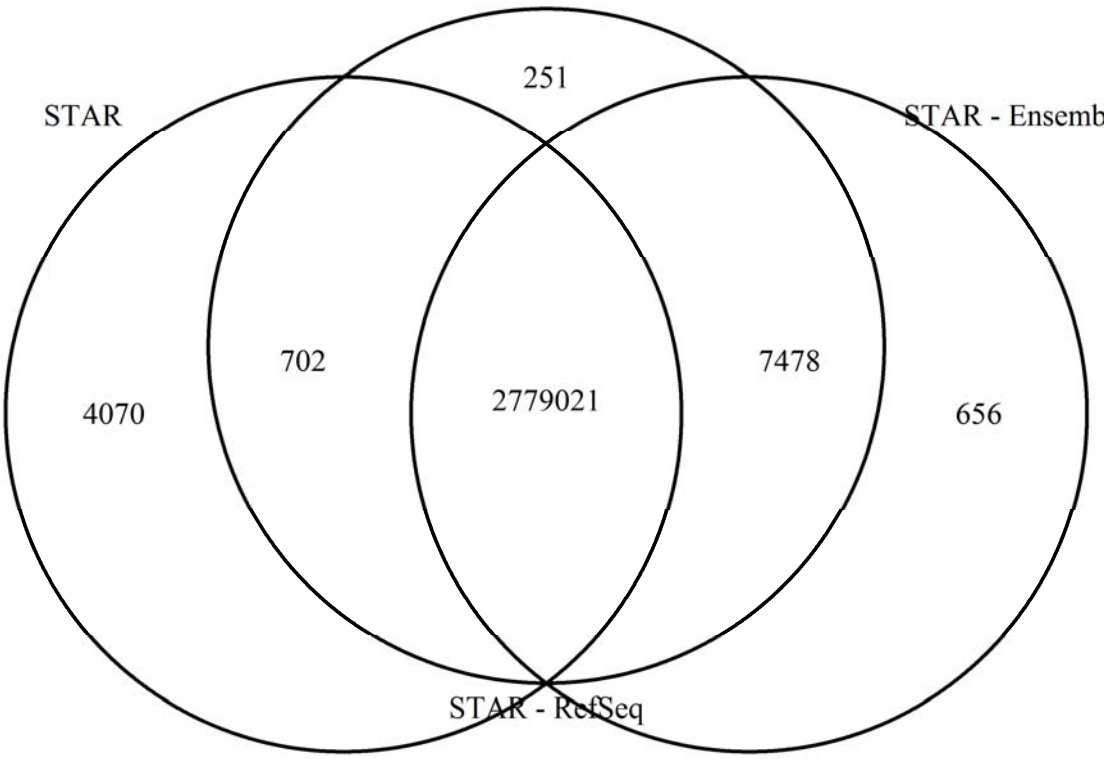


Figure 3.5: Nominal Changes to STAR Peaks by Choice of Annotation
Number of bases overlapping in peaks called by STAR using different annotations. STAR calls the same peaks for the most part, regardless of the annotation database used, only a few kilobases of peaks are unique to using an annotation database, which could aid in the alignment of spliced reads.

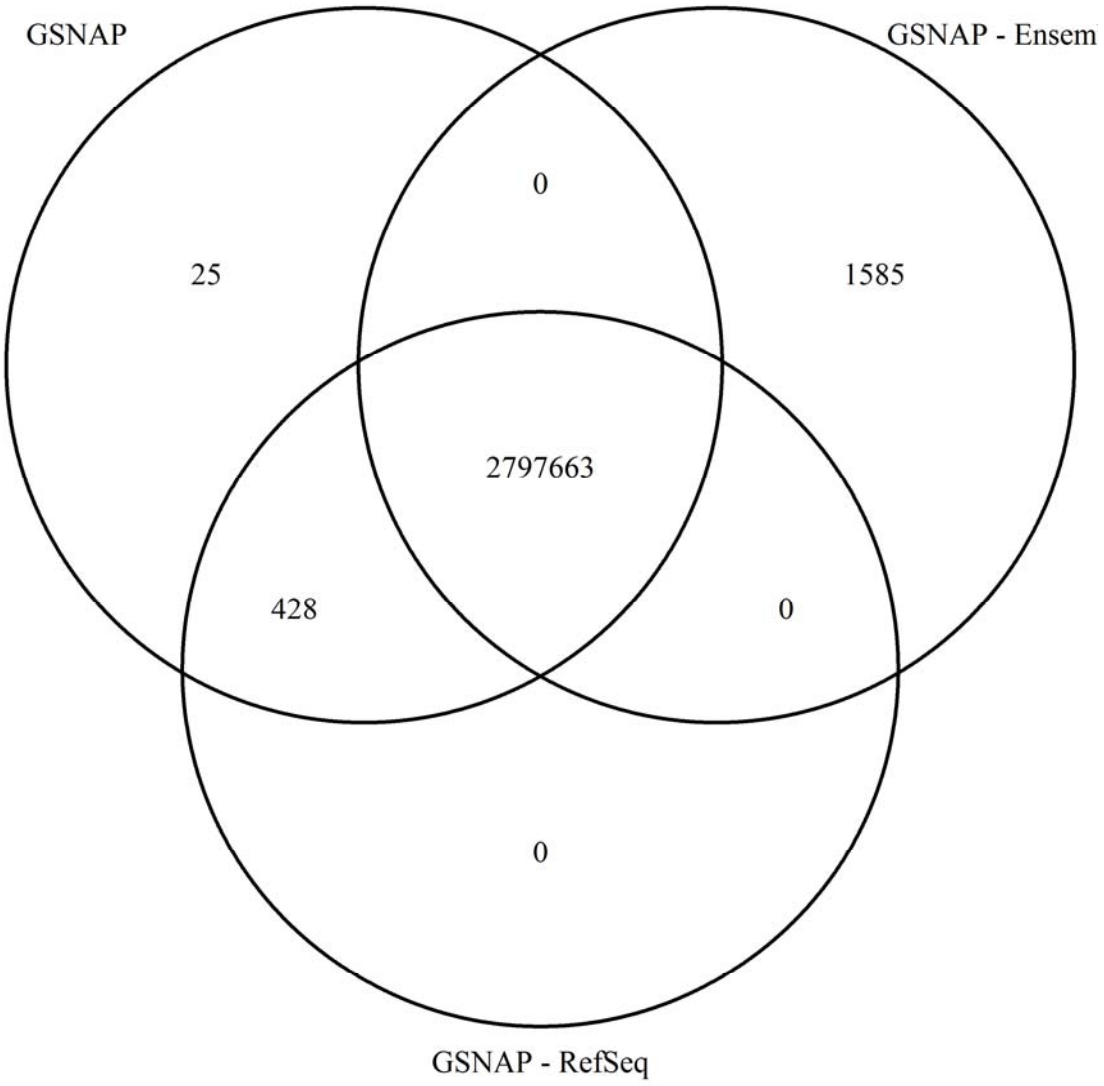


Figure 3.6: Very Small Changes in GSNAP Peaks Caused by Choice in Annotation

Number of bases overlapping in peaks called by GSNAP using different annotations. GSNAP has the best agreement between its aligner and usage of an annotation database. Despite using RefSeq or Ensembl annotations, the majority of the peaks are still called the same.

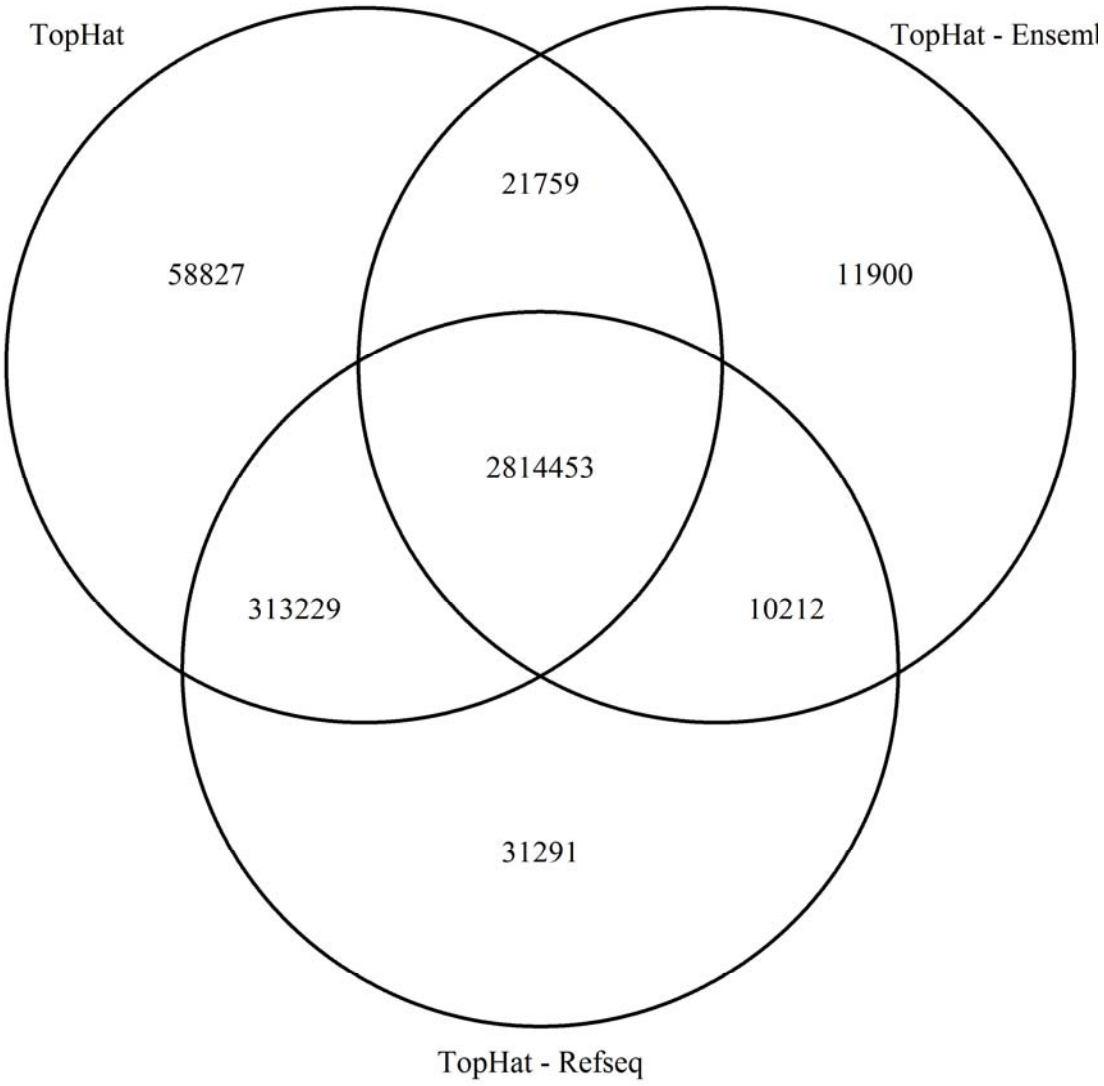


Figure 3.7: High Variation in TopHat Peaks Caused by Choice of Annotation
Number of bases overlapping in peaks called by TopHat using different annotations. TopHat shows the greatest variance with annotation databases. An empirical run of TopHat likely misaligns many reads, while the Ensembl alignment provided the best expected distribution of read counts earlier.

3.3.2 Fragment Shifting and Extension

In standard ChIP-seq protocols, DNA is fragmented to 200-300 base pairs using sonication, (Barski et al., 2007) but typically only the first 36-50 base pairs may actually be sequenced. This will result in two observed peaks, often on opposite strands, spanning a single chromatin binding site. The actual binding site can be found by calculating the fragment shift from paired-end sequencing data or estimating it from single-end sequencing data and shifting each read towards the 3' end accordingly, as used in MACS. (Zhang et al., 2008) The same can be said for RNA sequencing on the Illumina platform, where RNA samples are chemically fragmented to approximately 100 base pairs and usually sequenced single-ended 50 base pairs. Paired-end sequencing and longer reads may be used to achieve greater sequencing depth and additional splicing and isoform usage. (Li et al., 2014b) Newer Illumina strand-specific kits will result in reads on the 5' end of each RNA fragment.

MACS and exomePeak both use fragment shifting as a means of correcting the 5' library bias. (Meng et al., 2013; Zhang et al., 2008) In the context of DNA-sequencing and ChIP-seq peak calling, fragment shifting is a good solution to the 5' shift problem. The MACS method models this shift in single-ended data by plotting the Watson and Crick strands (positive and negative) independently and calculating the shift. However, using this method in RNA-Seq data is far less straightforward. First, the results from using MACS on the MeRIP-seq data from (Meyer et al., 2012) are shown in Figure 3.8.

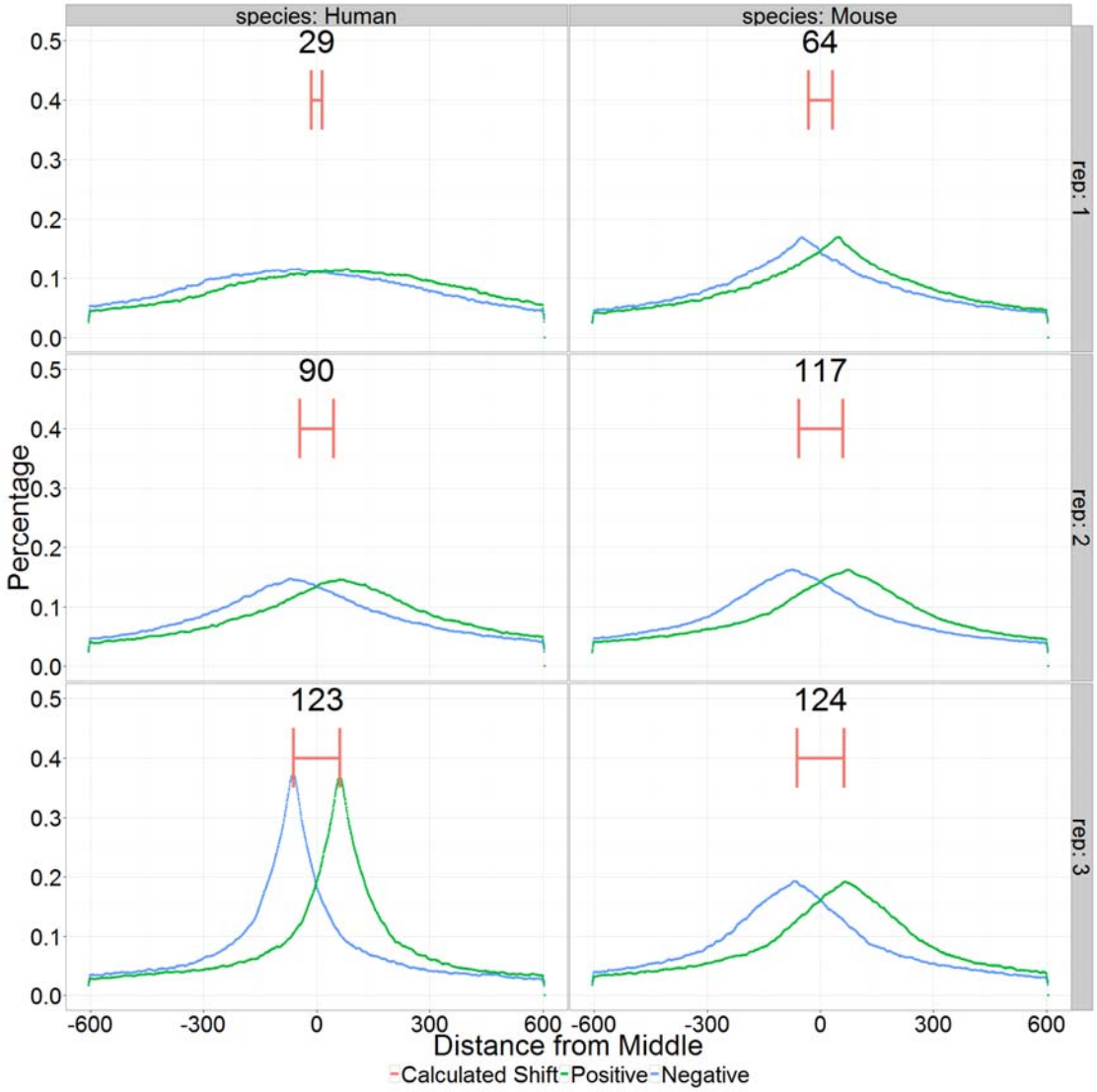


Figure 3.8: Fragment Shifts Computed using MACS2
The fragment shifts were computed for the original Meyer et al. (2012) data using MACS2 (Zhang et al., 2008), with the human samples on the left and the mouse samples on the right, and the three replicates by row. The red lines indicate the inferred fragment distance between the tags outputted by MACS2, and the blue indicates reads mapping to positive/Watson strand, the green to the negative/Crick strand.

Following standard MeRIP-Seq and Illumina sequencing preparation protocols, the fragment distributions should have been around 100 base pairs with only 50 base pairs sequenced on the ends, with the exception of the mouse sample

replicate 1 that had only 36 base pairs sequenced. The calculated shifts from MACS are inconsistent with the protocol and very greatly between replicates. This is likely because unlike ChIP-seq, which targets chromatin and transcription factors, m^6A and other RNA modifications are at single-base points. Multiple m^6A sites next to each other can confound this analysis and make it more challenging to correctly calculate this shift.

The software in exomePeak instead opts to use a user-supplied parameter to model this shift. Inputting a mean fragment length of 100 base pairs, the software shifts all reads half of that, or 50 base pairs, towards the 5' end. This circumvents the modeling challenges in MACS, but fragment shifting has its own problems in RNA-Seq. In the absence of large structural variation, shifting and extending DNA sequencing data is fairly trivial. In contrast, RNA-sequencing data from eukaryotes comes from a spliced transcriptome. Naively shifting reads in genomic space ignores read splicing and would result in shifting exonic reads into intronic spaces. exomePeak avoids this by shifting reads in the transcriptome space, but this then assumes that a particular read came from a specific annotated gene. The read could have come from a transcript that was alternatively spliced, contained a retained intron, or even from an immature transcript that had not yet been spliced. Making assumptions about the underlying data around these edge cases can lead to artifacts, as shown in Figure 3.9. Here, exomePeak incorrectly calls peaks in SLC9A3, a gene that is otherwise not expressed. By shifting the reads 50 base pairs to the 3' end, the algorithm shifts reads from BC013821, a gene that is on the opposite strand and part of which overlaps with an exon from SLC9A3, and incorrectly calls peaks in exons of SLC9A3.

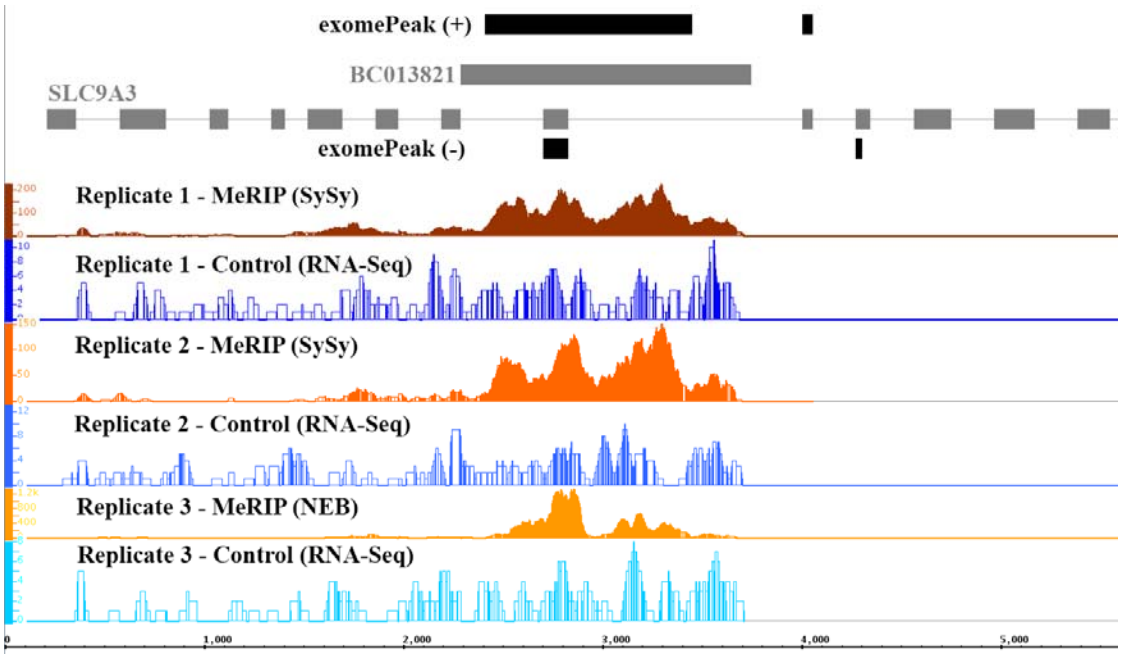


Figure 3.9: exomePeak Fragment Shifting Artifact
In black are exomePeak peaks depicted separately by strand and UCSC genes depicted in grey which were used to call the peaks in transcriptome space. Coverage from the RNA-seq replicates are shown, alternating MeRIP-seq and control-seq samples. exomePeak incorrectly calls peaks in SLC9A3 which is not expressed. Plotted using The Integrated Genome Browser (IGB). (Nicol et al., 2009)

Another potential solution to the 5' shift is fragment extension; extending each mapped read towards its 3' end until it is the average fragment length (100 base pairs), as used in ChIPseeqer's peak finder. (Giannopoulou and Elemento, 2011) Unfortunately, the same edges cases in fragment shifting are also present in extension in RNA sequencing data. The only advantage it has over fragment shifting is that many reads may in fact map over the m^6A site, which could lie anywhere along the 100 base pair fragment. Although MeRIP-seq data does show some 5' bias, there does not appear to be a clean solution to solve problem. Without paired-end sequencing data to confirm the exact length, splicing, and location of each fragment, the best solution is to simply not adjust

for it to avoid making assumptions about the underlying data and introducing artifacts.

3.3.3 Testing for IP Enrichment

After aligning to the genome, the next step is to test for statistically significant enrichment in the IP sample. MeRIP-Seq uses an antibody-based enrichment and antibodies are known to have non-specific binding, as well as the potential for other RNA fragments to be pulled down, resulting in background noise that can impede peak calling. MACS and exomePeak model enrichment using the Poisson distribution and edgeR uses the negative binomial distribution to model RNA-sequencing count data. (Meng et al., 2013; Robinson et al., 2010; Zhang et al., 2008)

The MeRIPPeR peak finding protocol uses Fisher's Exact Test, ensuring that the test only tests for significance in the direction of the IP. The original method utilized 25 base pair book-ended windows that spanned across genome, but this methodology can be expanded to include overlapping windows of a user-specified size. The Fisher's table used to calculate the p-value is shown in Table 3.1, which compares the enrichment in the current window to the enrichment observed outside of the window. Fisher's exact test is a non-parametric test that makes no assumptions of the distribution of the underlying data, and this particular table serves to normalize for differences in sequencing depth that may be present between the MeRIP and control samples. Fisher's exact test will return a p-value, which is then adjusted using Benjamini-Hochberg to account for the multiple testing problem.

Table 3.1: Fisher's Exact Table used to compute Fisher's Test

	# Reads in Window	*# Reads Outside Window*
MeRIP	[MeRIP Reads in Window]	[Total MeRIP Reads Mapped] - [MeRIP Reads in Window]
Control	[Control Reads in Window]	[Total Control Reads Mapped] - [Control Reads in Window]

3.4 Challenges in Peak Finding

Although the MeRIP-Seq protocol is fairly standard, many inconsistencies in the implementation of the protocol can introduces biases that complicate peak finding. Unlike eRRBS, which is chemically based, IP-based enrichment methods are highly sensitive to the antibody and its efficiency. In addition, the method of RNA isolation and choice of aligner all can have an impact on the ability to call peaks. Some of these challenges can be solved, while others, unfortunately, cannot be accounted for, but their effect on peak calling must nonetheless still be considered.

3.4.1 Antibody Non-Specific Binding

An antibody is a Y-shaped immunoglobulin (Ig) protein, traditionally produced by the immune system to identify bacteria, viruses, and other foreign agents. Most of the antibodies produced for MeRIP-Seq are created by injecting rabbits or mice with free m^6A in bovine serum albumin (BSA). The host produces antibodies in response to the m^6A and develops specific antibodies that bind to the m^6A antigen. These antibodies are then extracted and purified for use in immunoblots and IPs. Most of the antibodies used in MeRIP-Seq are polyclonal, meaning they are derived from different B cell lineages. Some monoclonal antibodies exist (Synaptic Systems #202 011 and #202 111 and others), but

their efficacy has not yet been tested in the IP. The purpose of the purification process is to isolate those antibodies that bind specifically to m^6A. The binding of the antibody to the target of interest is dependent on the specificity of the antibody to the epitope on the antigen. In the case of m^6A, the antibody must be able to distinguish between m^6A, adenosine, and other nucleotides. Regardless, non-specific binding, where the antibody binds to something other than the antigen, can still occur. Monoclonal antibodies have the advantage of higher specificity to the epitope, and in theory, should result in lower non-specific binding.

In addition, in the MeRIP-Seq protocol utilizes Dynabeads® M-280 Sheep anti-Rabbit IgG to pulldown the antibodies. The beads are first washed with BSA to reduce non-specific binding in IgG of the beads. The m^6A antibody and beads are then bound and washed to remove any antibodies that may not have bound. The RNA is then bound to the m^6A antibody and unbound RNA is washed. The immunoprecipitated RNA is ultimately extracted by using a magnet to separate the superparamagnetic beads bound to the sheep anti-rabbit IgG, which is bound to the m^6A sheep antibody, which itself is bound to the RNA. These multiple binding steps can not only affect the sensitivity of binding, but also introduce RNA fragments that do not contain m^6A sites. The m^6A-seq protocol utilizes Protein A instead of the Dynabeads, which can also introduce non-specific binding of its own, and thus the protocol recommends using a beads-only control to measure the level of background binding. (Dominissini et al., 2013)

IP-enrichment methods rely on both the successful pulldown of RNA fragments that contain m^6A sites, as well as the removal (through washing) of RNA

fragments that do not. The (Meyer et al., 2012) paper showed using real-time polymerase chain reaction (RT-PCR) that a single round of IP achieved 70-fold enrichment of m^6A fragments over background binding and two-rounds achieved 130-fold enrichment. These results represent the enrichment at the fragment level, but after library preparation, the sequencing results are likely more varied and dependent on each transcript and far more features. Unfortunately, without good negative controls or the ability to verify the lack of m^6A sites, the exact amount of non-specific binding is difficult to measure.

Spike-in RNA-sequences have often been used to assess quality control metrics of RNA-sequencing data. (SEQC MACQ-III Consortium, 2014) RNA oligonucleotides that are designed to not map to any known sequences, these sequences can be "spiked-in," hence the name, to measure library preparation and sequencing biases. For the purposes of MeRIP-seq, four such sequences were constructed using in-vitro transcription (IVT) based on DNA oligonucleotides of lengths 71-100. The oligonucleotides were specifically designed with only a single thymine site, resulting in the creation of a single adenosine in the corresponding RNA transcript. This enables running two simultaneous IVT reactions, one that utilizes only adenosine in its mix of nucleoside triphosphates (NTPs) and another that contains only m^6A. The two reactions can then be normalized and mixed at specific ratios, yielding spike-ins that have varying percentages of m^6A. These spike-ins were synthesized by Kate Meyer, PhD and sequenced as part of an exploratory m^6A project. The log 2 peak enrichment and the corresponding percentages of m^6A used are shown in Figure 3.10. Peaks were not called on the 0% spike-in, which did not pass Fisher's exact test in all replicates, but some replicates still show a high degree of non-specific binding. Nonetheless, the somewhat linear trend in the increase

in observed log 2 peak enrichment does show promise in the ability to recapitulate m^{6}A levels with the IP.

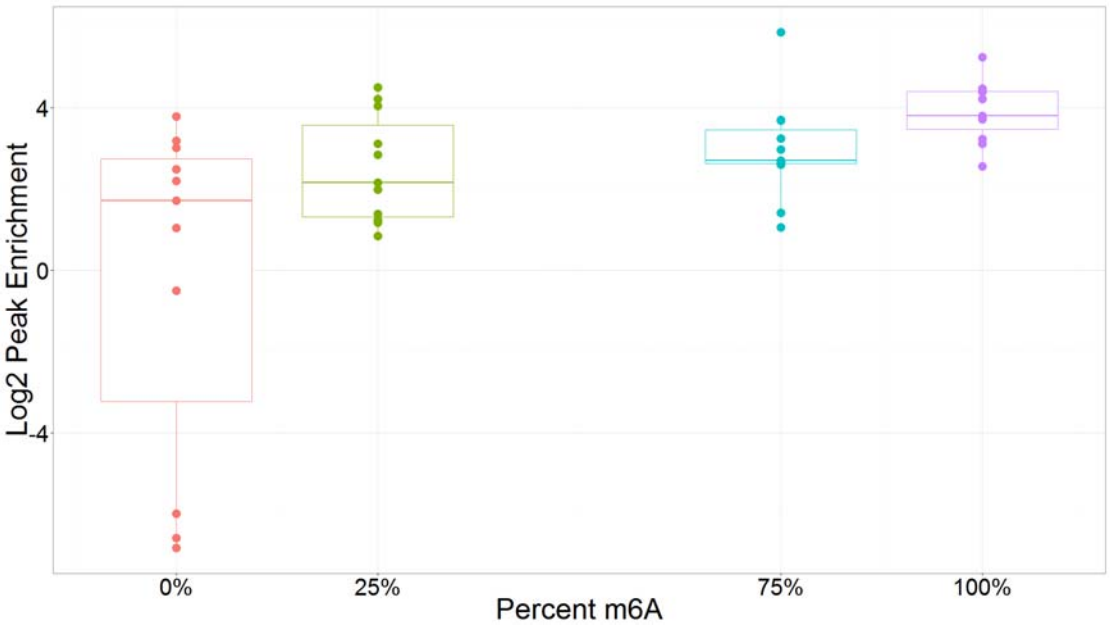


Figure 3.10: Linear Distribution of Spike-In RNAs Correlates with Methylation Fraction
The distribution of $\log_2$ peak enrichments of four spike-ins used in an experiment are shown, with the percentage of methylation on the x-axis and the distribution of $\log_2$ peak enrichment on the y-axis. The percentage of m^{6}A present compared to the enrichment shows a high degree of variability in the IP.

3.4.2 MeRIP-Seq IP Enrichment

In the previous section, the properties of the antibody and its effect on specific and non-specific binding was discussed. In the (Meyer et al., 2012) paper, the IP enrichment was defined as the fold-enrichment of m^{6}A containing fragments over non-specific binding. Ideally, the enrichment of a single peak is a function of this enrichment, the expression level of that transcript, and the percentage of transcripts that contain m^{6}A in the peak region. Unfortunately, the IP enrichment itself is a function of multiple factors, including the amount of RNA used in the

IP, and is potentially variable with respect to each RNA fragment and influenced during PCR by GC-content biases and the mappability of the transcript by the aligner. (Li et al., 2014a; Li et al., 2014b; SEQC MACQ-III Consortium, 2014) The input titration experiment and multiple rounds of IP initially used to determine the input levels of MeRIP-seq, discussed earlier in section 2.3.2 Input Requirements, can be used to elucidate the variability of IP enrichment across replicates, and the sequencing impact of utilizing two-versus-one round of IP. Using the default MeRIPPeR peak caller, the number of bases in peaks called is shown in Figure 3.11. Unfortunately, some of the replicates did not perform as well as expected, such as the first replicate of the single-round 300 microgram IP and the second replicate of the two-round 100 microgram IP.

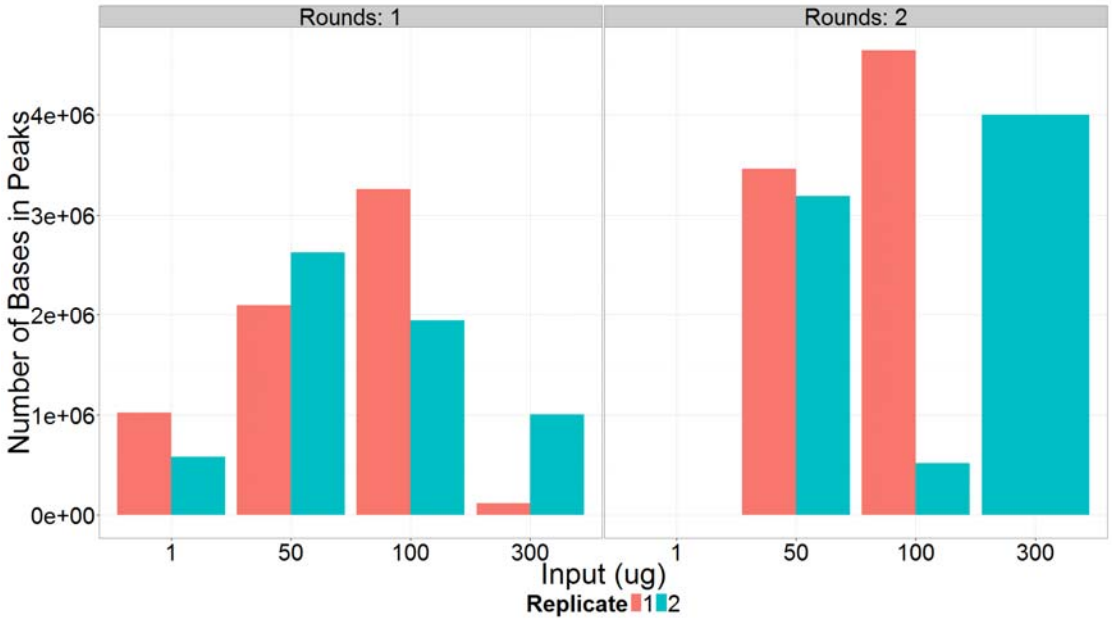


Figure 3.11: Increased Number of Peak Bases in 2-Round IPs
The number of bases in peaks is shown, with one-round IPs shown on the left, two-round IPs on the right, and the input in micrograms on the x-axis. Some of the replicates did not perform as well, and are likely a technical failure in the IP and not specific to the input amount.

Plotting the number of bases spanned by peaks, versus total number of peaks, is a more fair representation of the genomic span of the peaks. Using solely the number of peaks, for example, does not show the size of each peak. Using two-rounds of IP results in more peak bases being called, as expected, with reduced background noise from non-specific binding. Plotting the distribution of peak enrichments using sliding windows of size 100 and step-size 25 across the union of all peaks in Figure 3.12 confirms a strong lack of enrichment in those replicates, as well as strong disparities in the distribution of peak enrichments in the one-round 300 microgram IP. Although the first replicate appears to be a cleaner IP, the fewer number of peak bases called in Figure 3.11 demonstrate that the first replicate had poorer IP efficiency or higher specificity. The failed replicates are likely technical artifacts from failed IPs and will be excluded as outliers.

Removing the peaks from the failed replicates and recalculating the log 2 peak window enrichments using the same sliding window method in Figure 3.12 shows a clear increase in the average peak enrichment in the two-round IPs over the one-round IPs, as expected, shown below in Figure 3.13. It should be noted that in each IP, there are clear windows with log 2 peak enrichments below zero, which are from peaks not called in that particular replicate but in other replicates, and represents technical variation in the IP, because the initial RNA pool used for this experiment was the same. With both replicates successful in the one-round and two-round 50 microgram experiments, the increase in IP efficiency can be computed by comparing the mean enrichments observed. Using 100 base pair sliding windows at a step size of 25 base pairs across the intersection and union of each set of peaks, the one-round and two-round IPs separately, the enrichment scores are plotted in Figure 3.14.

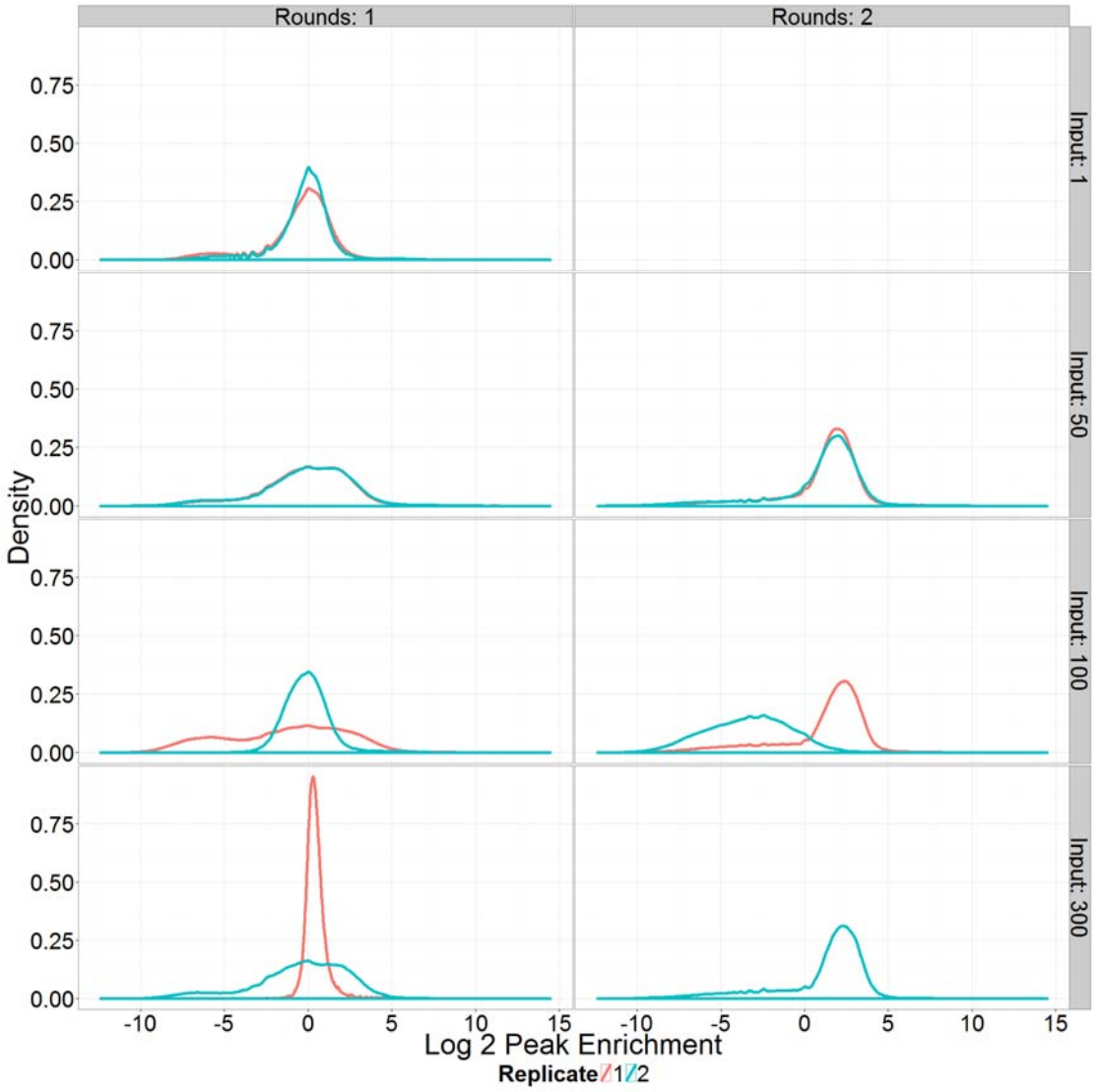


Figure 3.12: Density of Peak Enrichment Windows in IP Input Test
Density plot of log_2 peak enrichment, with one-round IPs shown on left and two-round on the right, input in micrograms varying vertically, and replicates by color. The density distributions confirm lack of enrichments in technical replicates, showing far less enrichment in the first replicate of the 100 microgram one-round IP and the second replicate of the two-round 100 microgram IP. The 300 microgram one-round IP replicates show a strong disparity in their enrichment densities.

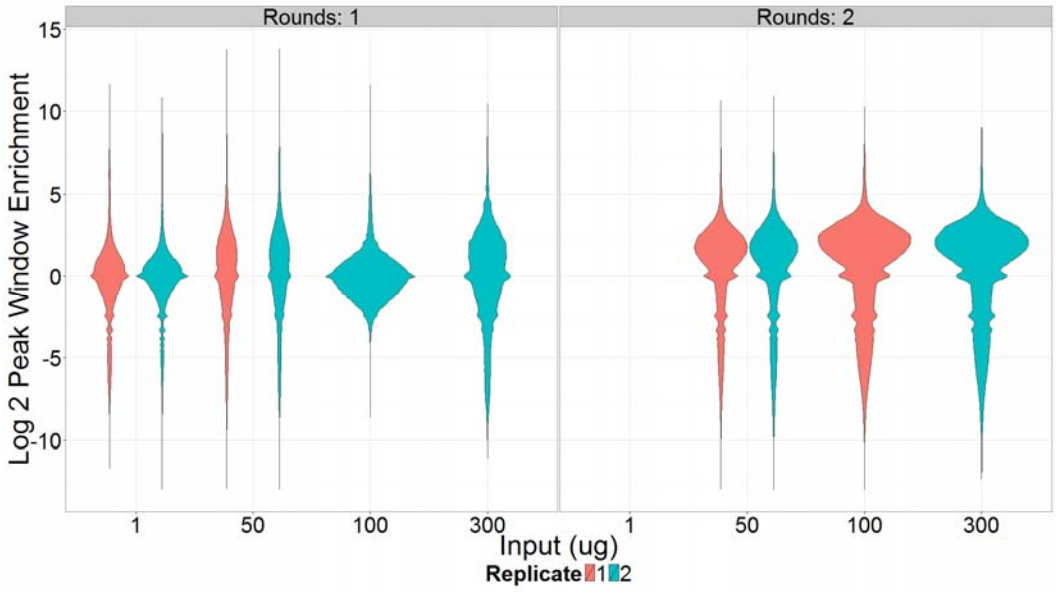


Figure 3.13: Increased Mean Peak Enrichment in 2-Round IPs
Violin plot of the peak enrichment from Figure 3.12 with one-round IPs on the left, two-round IPs on the right, input in micrograms on x-axis, and replicates by shading, shows an increase in enrichment of two rounds of IP over one.

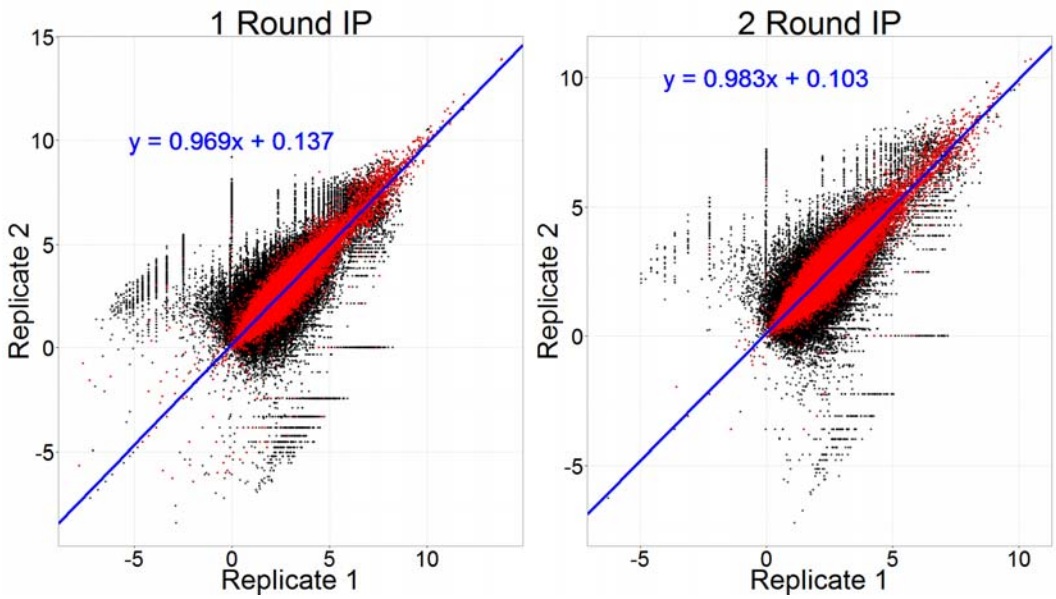


Figure 3.14: Linear Correlation of Technical Replicates in IP
Log_2 peak enrichment scores with replicate 1 on x-axis and replicate 2 on y-axis, windows from the union of peaks in black and intersection in red, and linear fit shown in blue, one-round IPs shown on left and two-round on right. The enrichment scores show a strong correlation between replicates.

The details of fitting the linear model to the CPMs are summarized in Table 3.2. The linear fit R^2 values for the single-round and double-round IPs are 0.8911 and 0.8690, respectively. The linear model was fit to the windows from the intersection of the peak replicates. A linear model fit to the windows from the union of the peak replicates resulted in a fit skewed from observable trend and the intersection fit. The union fit was skewed by windows with very low to near zero read counts in one of the replicates and very low read counts in the other. This is consistent with previous findings in RNA-seq of being able to make more confident estimations of RNA transcript abundances of genes that are highly expressed over those that are not. (Anders and Huber, 2010) The Pearson correlations of the union window CPMs are 0.5642 and 0.7628, respectively, which is well below previously reported correlations for technical replicates of RNA-seq data, (Yamamoto et al., 2014). The same Pearson correlations of the intersection window CPMs are 0.9440 and 0.9322, which shows a very strong correlation of peak enrichments in peaks common to technical replicates.

Table 3.2 Summary of 50 Microgram IP Linear Modeling in Replicates

	1-Round IP	2-Round IP
Linear Modeling		
Intercept	0.1366	0.1034
Slope	0.9692	0.9830
R^2	0.8911	0.8690
P-value	< 2.2e-16	< 2.2e-16
Union Window Correlations		
Pearson Correlation	0.5642	0.7628
Variance	1.4879	0.8913
Intersection Window Correlations		
Pearson Correlation	0.9440	0.9322
Variance	1.6475	0.8057

3.4.3 The Advantages of Two Rounds of IP

The two-round IP is implemented by first performing a single round of IP and then inputting that RNA back into a second round of IP. The experimental design was constructed such that the two-round IP replicates would be matched to the one-round IP replicates, in that their input was the same pool of RNA that was sequenced as the one-round IP, as discussed earlier in Figure 2.3. The results of comparing the two-round IPs to the one-round IPs using their enrichment scores are depicted below in Figure 3.15. There's a clear subset of peaks that are present in the one-round and not in the two-round IP, and vice versa. However, the results from the Figure 3.12 and Figure 3.13 earlier do show that the two-round IPs achieve better IP enrichment and better technical replication, with lower observed variance in peak enrichment.

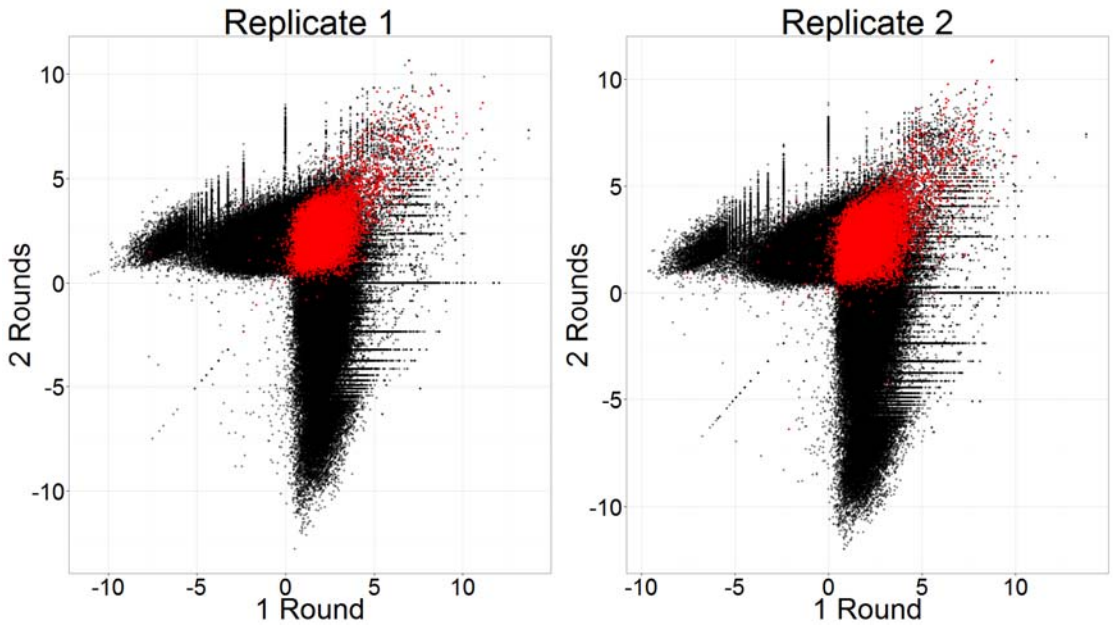


Figure 3.15: High Variability in Peak Enrichment Between 1- and 2-Round IPs
The $\log_2$ peak enrichment for windows shown for 1-round IPs on x-axis, 2-round on y-axis, with the first replicate on left and second replicate on the right, windows from union of peaks in black and intersection of peaks in red. A large number of peaks are found unique to each set of IPs.

3.4.4 Determining and Correcting Batch Effects

Multiple batch effects were discussed, including the input amount of RNA, the IP efficiency, and the usage of one versus two rounds of IP. The IP efficiency is itself likely dependent on the input amount of RNA and other factors. Using principal component analysis on the peak enrichment scores, the dominant feature is the rounds of IP, shown in Figure 3.16. The second dimension could correspond to the IP input, if ignoring the replicates that showed poorer IP efficiency. Figure 3.17 shows a scree plot, or the variance, in the PCA analysis, demonstrating that the first two components capture the majority of the variance.

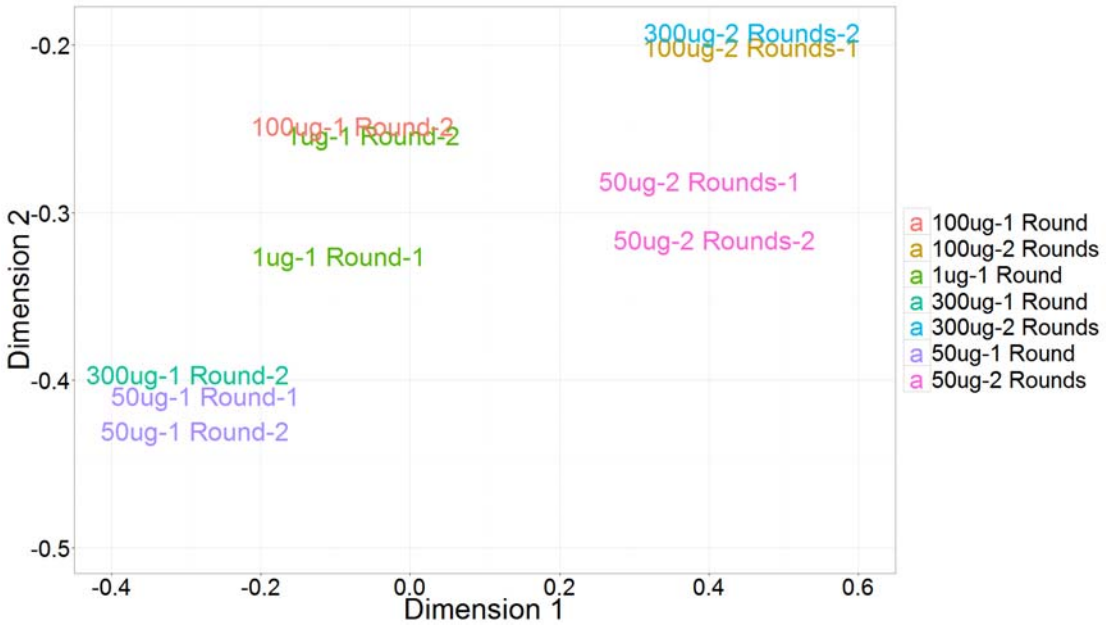


Figure 3.16: Greatest Separation in IP Input Test Corresponds to Rounds IP
A principal component analysis (PCA) of the IP Input Test IP enrichments shows the first dimension corresponding to the rounds of IP. IP replicates with low technical variability cluster well together. Samples are colored by input and rounds of IP.

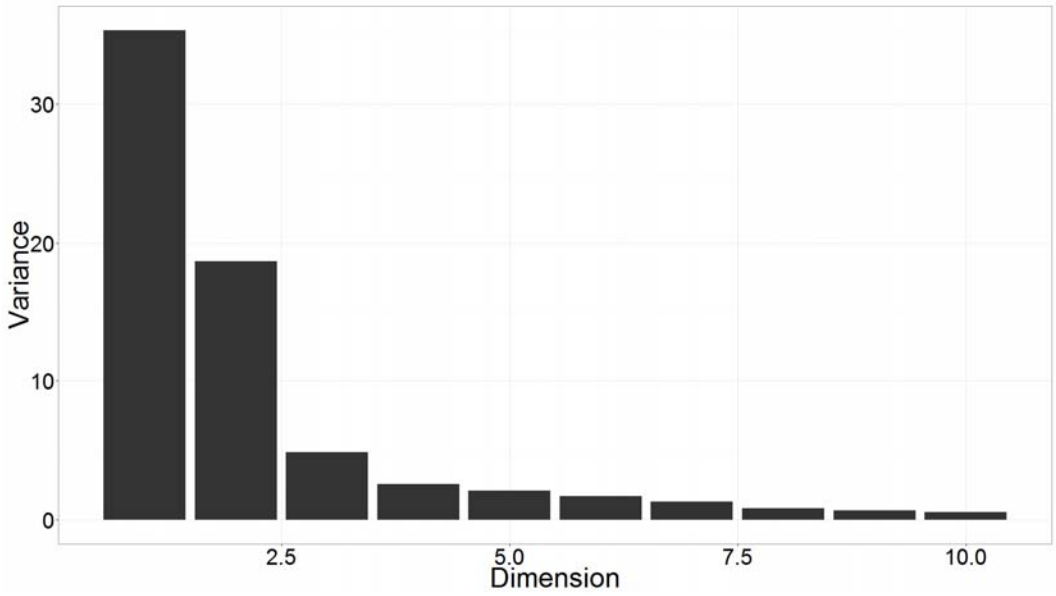


Figure 3.17: First Two Dimensions Capture Majority of Variance in IP Input Test PCA Analysis
A scree plot showing the variance of the principal component analysis of the IP input test peak enrichment scores.

In computing the peak enrichments, the raw counts are converted to counts per million, normalizing by the number of mapped reads to account for differences in sequencing depth. The trimmed-mean method (TMM) (Robinson and Oshlack, 2010) from edgeR (Robinson et al., 2010) is typically used to adjust for the number of genes in the sample, used to scale the effective library size. This scaling factor can be applied to the MeRIP-Seq samples to account for the amount of m^6A, and the non-specifically bound fragments, present in each of the replicates. Figure 3.18 shows the TMM scaling factors for the MeRIP-seq samples, which shows the two-round IPs have often double the scaling factor of the single-round IPs. A larger scaling factor, applied in the denominator, would result in further shrinkage of the counts, accounting for the increased enrichment observed. This scaling factor can be used to effectively account for various batch effects, especially the number of rounds of IP in the sample. PCA analysis of the scaled enrichments, shown in Figure 3.19, does not show as clear separation on the first or second axes corresponding to the rounds of IP.

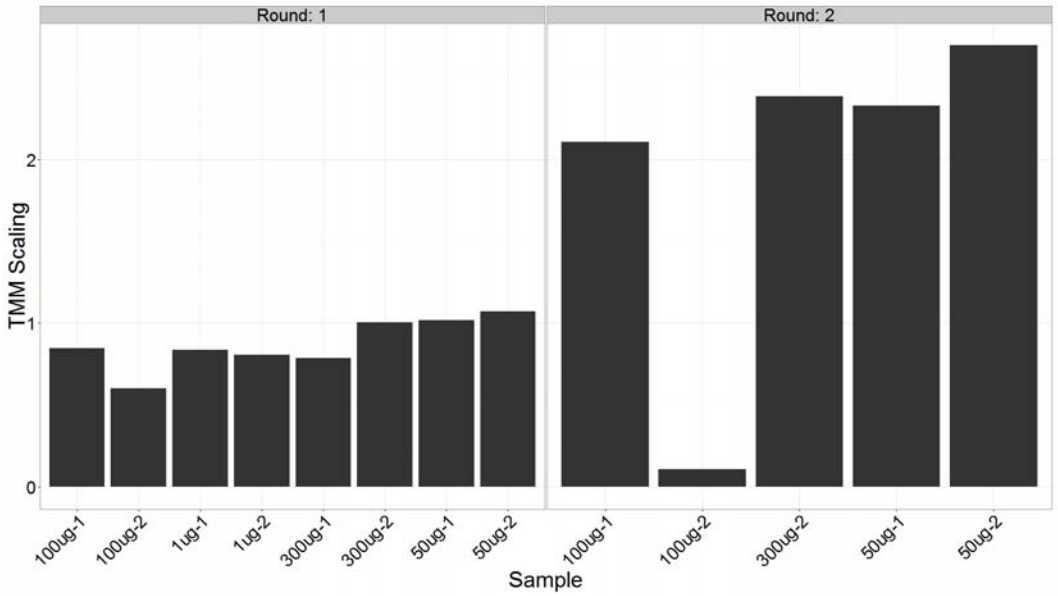


Figure 3.18: Double TMM-Scaling Factors in 2-Round vs 1-Round IPs
The TMM scaling factors from edgeR computed only on the MeRIP-Seq samples (y-axis), with input in micrograms and replicate on x-axis and rounds of IP separated with one-round on left and two-round on right. The two-round IPs shows a clear increase, approximately double, in the scaling factor applied, relative to the single-round, with the exception of the failed replicate.

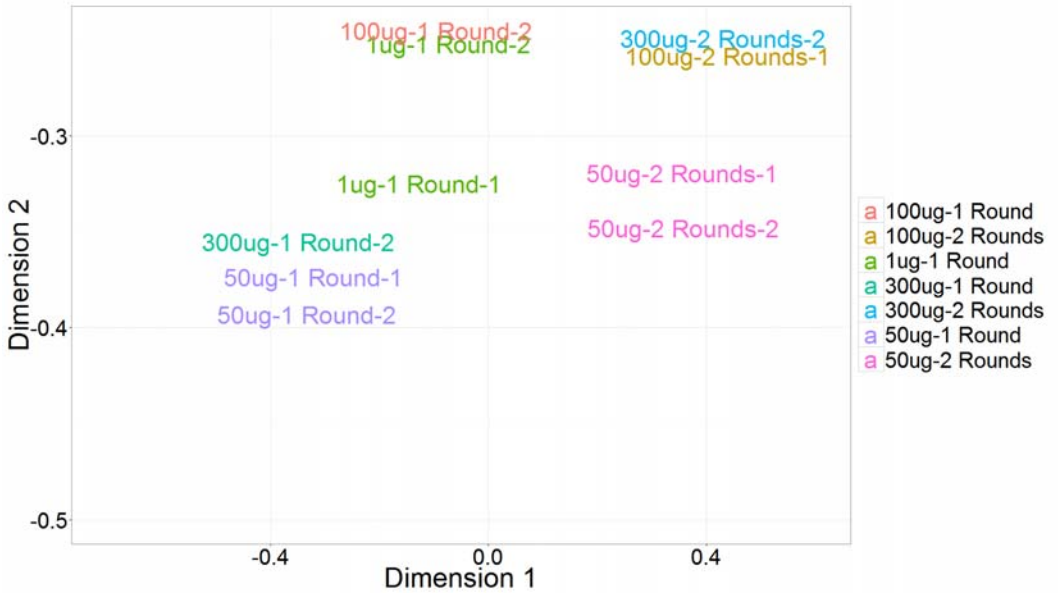


Figure 3.19: TMM-Adjusted PCA Shows Better Clustering of Samples
Applying the TMM scaling factors to compute the counts per million (CPM) prior to calculating normalizes for various batch effects, including rounds of IP. Samples are colored by input and rounds of IP.

3.4.5 Ribosomal RNA Contamination

Earlier, the sequencing consequences of rRNA contamination was discussed. During peak analysis, rRNA sequences can also have a profound impact on the peak calling. In the trimmed-mean method (TMM) (Robinson and Oshlack, 2010) paper, the abstract example the authors depicted described two sample sets, samples A and B, which had been sequenced with the same number of reads each. Sample A has twice as many genes expressed as Sample B, and the coverage of each gene in A is therefore half as that in B. Traditional library depth normalization methods use the total number of mapped reads, which in this case is the same for both samples, and the majority of genes between the two samples would incorrectly be calculated as differentially expressed.

The rRNA contamination in MeRIP-Seq can be viewed similarly as an unknown set of genes that consume sequencing reads. Without properly normalizing for this contamination, the coverage in each genomic window is computed to be far lower than it actually is. Furthermore, since the amount of rRNA contamination is far greater in the control RNA-Seq samples than the MeRIP samples, the effect is observed to a much higher degree in the control samples. The end result is that the coverage is estimated to be lower in the control than it should be, relative to the IP, and more peaks are called, artificially and inadvertently increasing the false discovery rate (FDR). Insufficient rRNA depletion can occur for a variety of reasons, from poor ribosomal depletion (Meyer et al., 2012) to skipping the step in its entirety (Dominissini et al., 2012).

In the event that rRNAs were poorly depleted, one potential solution is to computationally remove the contaminating reads. That is, the reads can be first aligned to known ribosomal sequences, and only those reads that do not align

are kept for subsequent analysis. In contrast to Figure 2.1, Figure 3.20 shows the dramatic reduction in rRNA contamination after *in silico* removal.

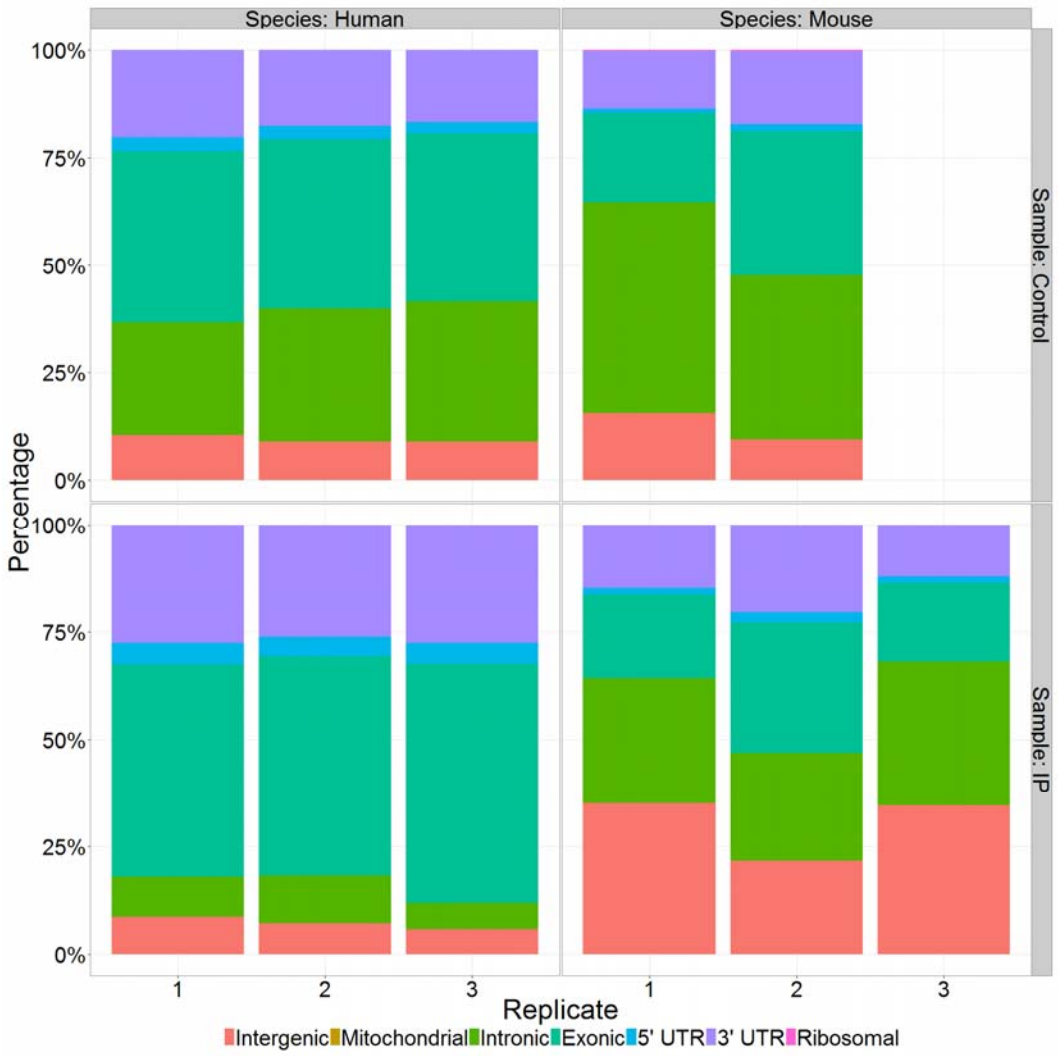


Figure 3.20 Successful *In Silico* Removal of rRNA Contamination in Meyer et al. (2012) samples.
Percentage of reads mapping to gene features shown on y-axis, with sample type and replicate on x-axis. Gene features are colored, with intergenic in salmon, mitochondrial in dark yellow, intronic in green, exonic in teal, 5' UTR in cyan, 3' UTR in purple, and ribosomal in pink. Aligning the reads first to a ribosomal RNA reference and then aligning only those reads that did not map to rRNA regions results in a highly successful removal of rRNA contamination.

The caveat here is that when the original sample already has low sequencing coverage, filtering out reads can result in very few mapped reads remaining, which may not provide enough coverage to call peaks, as shown in Figure 3.21.

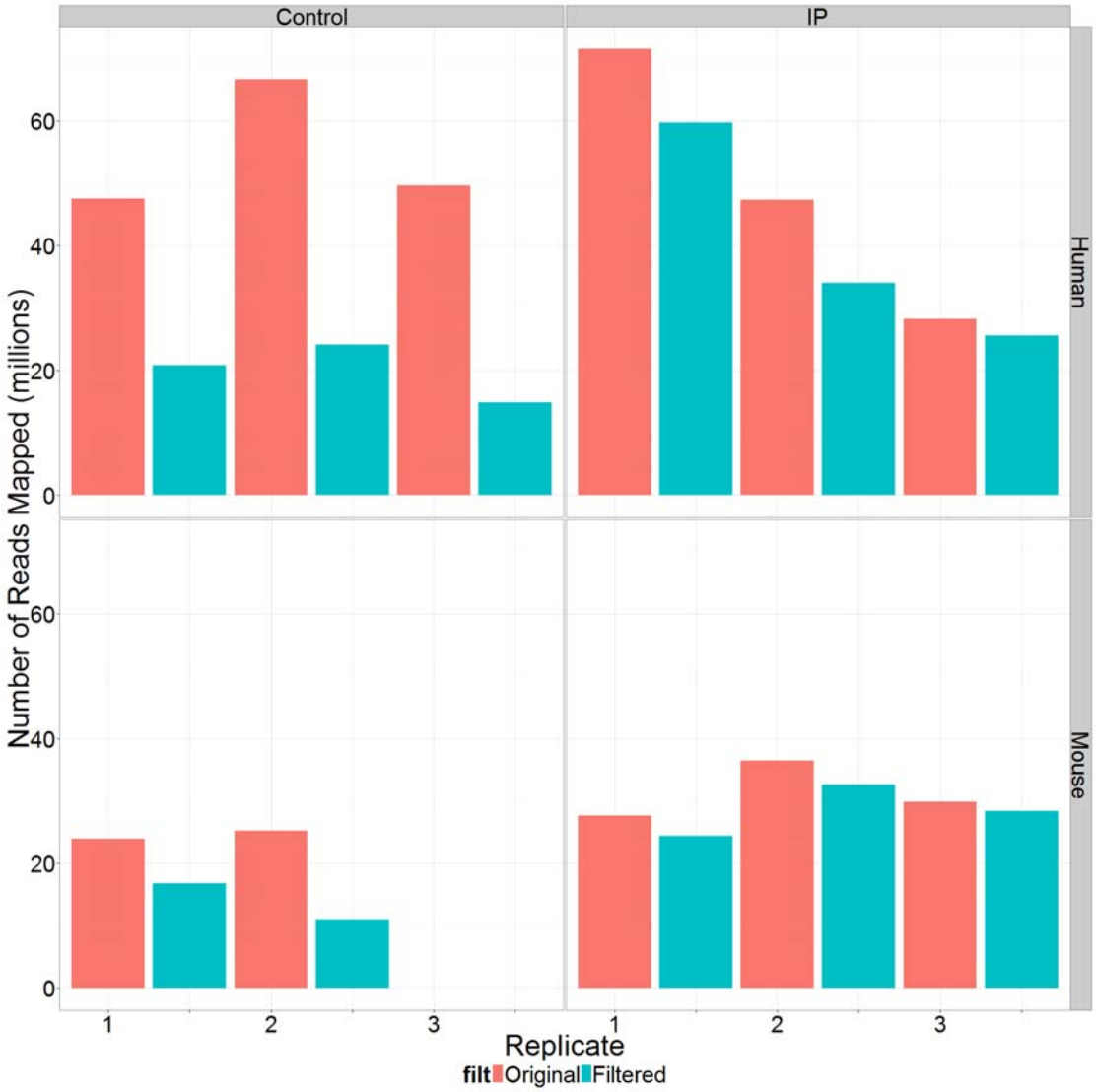


Figure 3.21 Loss in Total Number of Reads Mapped Following rRNA Removal
The number of reads mapped (y-axis) before (salmon) and after (cyan) *in silico* filtering, with control samples on left, IP samples on right, human samples on top, mouse samples on bottom, and replicate varying across x-axis. *In silico* removal of rRNA reads when a high degree of rRNA contamination is present results in a dramatic reduction in the total number of reads mapped post-filtering.

3.4.6 Splice Junctions

Splice junctions in the transcriptome present a challenge both in the alignment, as well as in peak calling. The advantage of exomePeak is that it can call peaks that may span across a splice junction by choosing to operate in the transcriptome space. This comes at the cost of being restricted to an annotated set of exons and prevents analysis of intronic and novel unannotated regions. MeRIPPeR's original genome-based window method successfully identified peaks within exons, but windows that span an exon-intron boundary may lose sensitivity due to mapping issues. For example, in Figure 3.22, the default MeRIPPeR genomic-based windows falls short at the end of the exon, likely because the next 5' window falls mostly in intronic space with far fewer reads.

One solution is to augment the traditional genomic based-windows with windows spliced across exon-exon splice junctions, shown in Figure 3.23. These augmented windows can be inputted from an annotation set, such as RefSeq or Ensembl, or from empirical splicing data output from STAR or TopHat. These windows specifically avoid the mapping issues at the exon-intron boundary, and take into consideration spliced-reads.

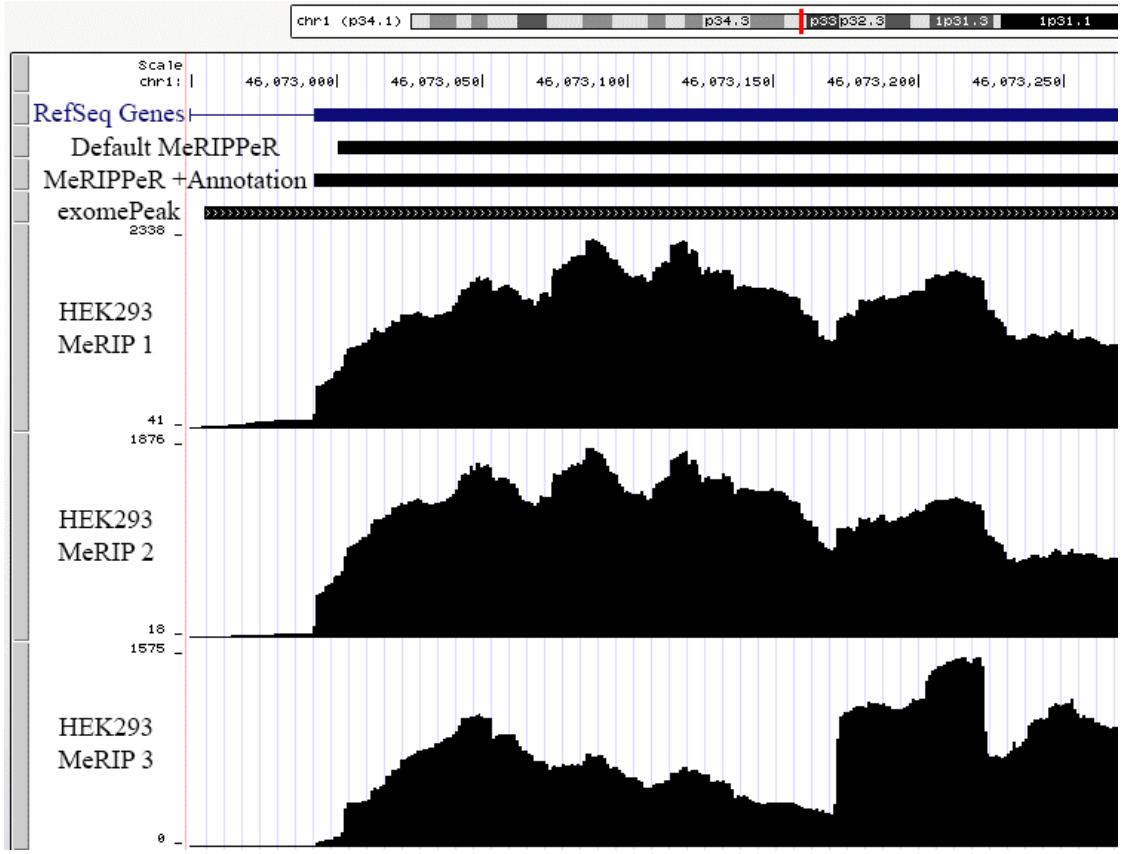


Figure 3.22 Annotation Supplement Adds Increased Coverage at Exon Ends
UCSC Genome Browser showing coverage and changes in peaks after augmenting spliced windows. Supplementing the traditional MeRIPPeR genomic windows with spliced peaks can help extend peaks to the exon edge.

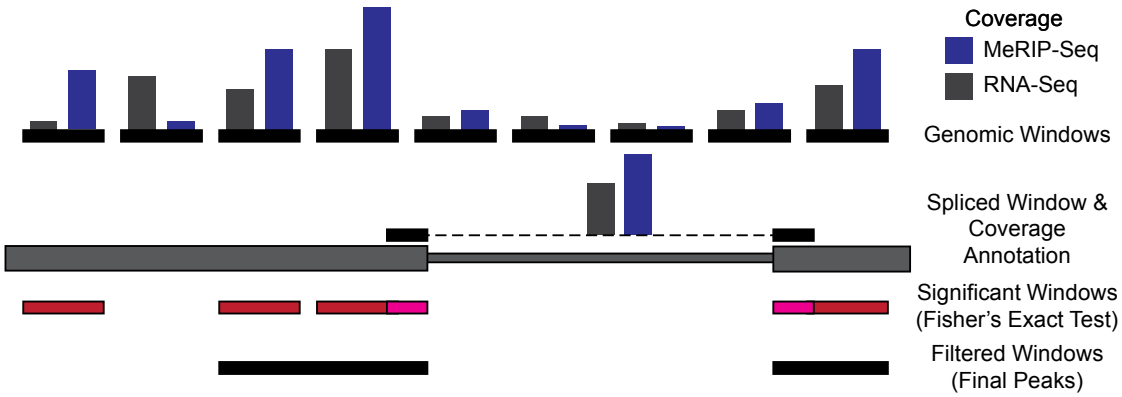


Figure 3.23 Augmented MeRIPPeR Window Method
The traditional genomic windows are shown in black with an abstract example. The addition of spliced windows is shown in the middle and how these spliced windows augment the genomic peaks, with the final peaks on the lowest level.

3.5 Comparison with Existing Peak Callers

Using the human data from (Meyer et al., 2012), the peak callers MeRIPPeR, exomePeak, and MACS can be compared. Without ground truth to characterize the false positives and negatives in each peak caller, the qualitative features of the peaks can be compared. All of the peak callers call roughly the same number of peaks, as shown below in Figure 3.24, though MeRIPPeR calls the most peaks. As was discussed earlier, the raw number of peaks does not characterize the sizes of each peak, but a weighted Venn diagram of peak regions, weighted by peak size, in Figure 3.25 shows MeRIPPeR still calls the most peak regions relative to the other peak callers. Despite these characteristics, the global metagene distribution of peaks remains relatively unchanged, as shown in Figure 3.26.

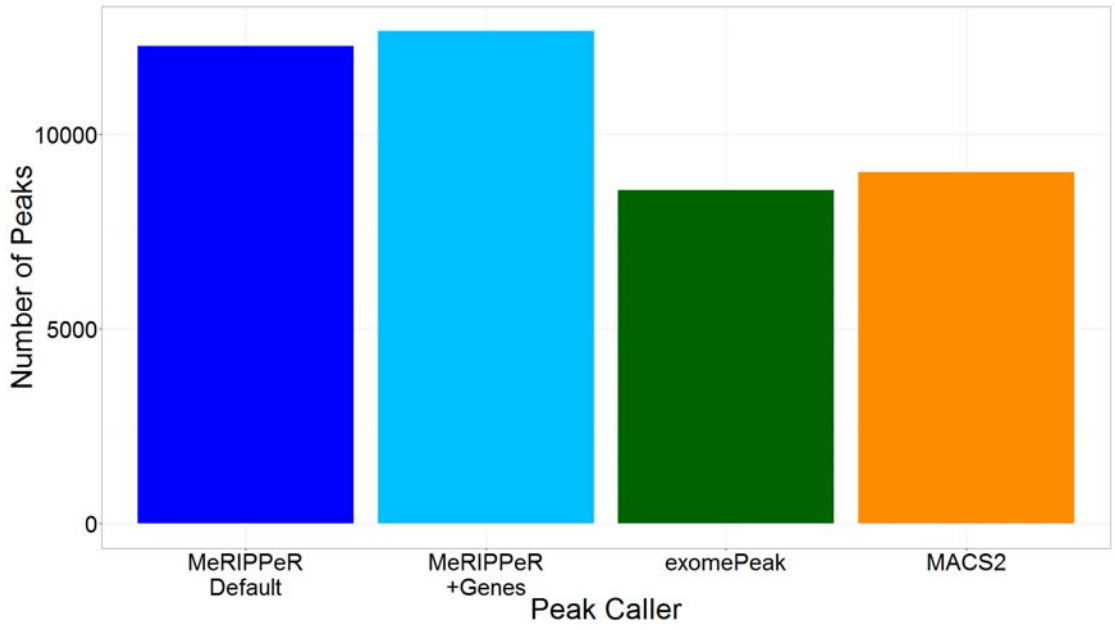


Figure 3.24 Number of Peaks Called by Different Peak Callers
Number of peaks called (y-axis) by each peak caller (x-axis). MeRIPPeR calls the most number of peaks, though the characteristics of each peak can be very different in the human HEK293 data.

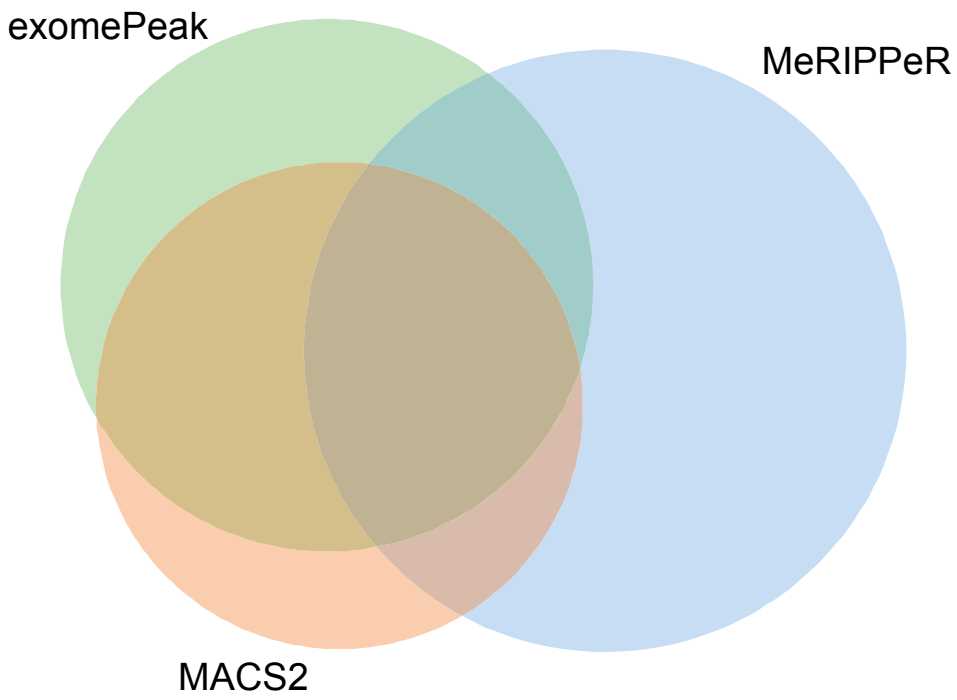


Figure 3.25 MeRIPPeR Calls Unique Set of Peaks
A Venn-diagram weighted by the number of bases called in each peak still shows MeRIPPeR calls the most peaks in the Human HEK293 data.

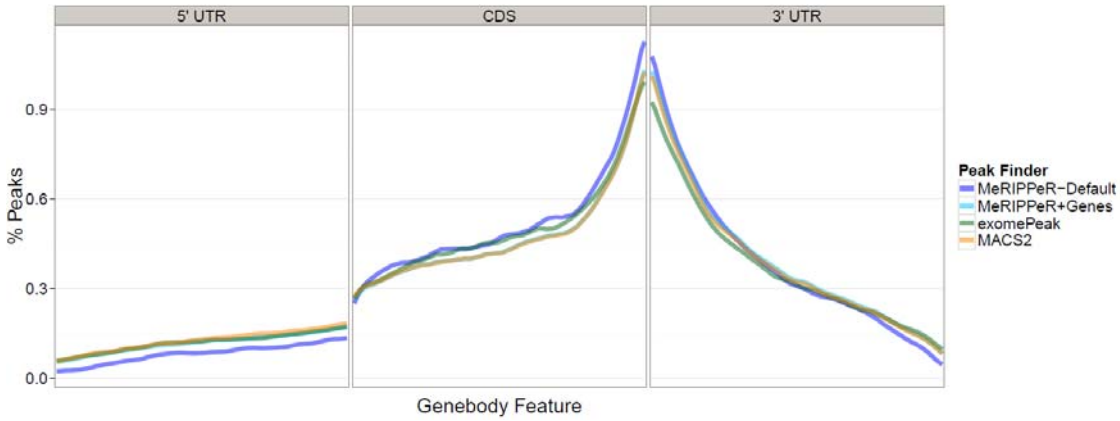


Figure 3.26 Metagene Comparison of Peak Callers
Metagene distribution showing binned 5' UTR, CDS, and 3' UTR for MeRIPPeR-Default (blue), MeRIPPeR+Genes (cyan), exomePeak (dark green), and MACS 2 (orange). The metagene distribution shows the peak callers not only recapitulate the stop codon enrichment of m^6A sites, but that they in part produce the same global m^6A signature.

The advantage of exomePeak, the authors claim, is its ability to better call peaks across splice-junctions by operating in the exonic space. The original MeRIPPeR algorithm was built with genomic-based windows, which could miss peak regions that may be spliced across an exon-exon splice junction, as discussed earlier in 3.4.6 Splice Junctions. The augmented splice-junction windows serve to solve that problem, and the method shows that far more spliced-peaks are called when using the augmented windows, shown in Figure 3.27, over the traditional MeRIPPeR algorithm. exomePeak calls the most spliced-peaks, and it has the advantage that it reports the peaks as spliced in the bed 12 file format, but at the cost of only being able to interrogate exonic regions. Figure 3.28 shows the number of bases in peaks mapping to intronic and intergenic regions. These peaks may come from immature transcripts, novel isoforms and genes, or from retained introns, none of which can be interrogated by using the other peak callers.

The foremost advantage of MeRIPPeR is its implementation in Java, built on the *htsjdk* package, as shown in Figure 3.29. The single-core MeRIPPeR performs the fastest read coverage calculations and computes Fisher's exact test very fast with its own internal implementation. MeRIPPeR was also designed with multi-core server clusters in mind, and the usage of multiple cores enables even faster throughput. exomePeak unfortunately computes the coverage in R, which results in a very slow implementation, taking nearly 4.5 hours to call peaks on the data set, longer than most aligners would have taken to align the same data to the genome.

The authors of exomePeak claim that their advantage is examining peaks in the transcriptome space, leading to clear identification of and classification of

peaks. They further purport that this reduces ambiguity when dealing with multiple transcript isoforms for a single gene and that they can correctly identify spliced peaks as a single peak. While directly calculating peaks in the transcriptome space does ensure clear gene assignment and spliced peaks, it tends to hide the problems they mention, rather than solving them. For example, peaks called by exomePeak often overlap with one another, due to overlaps in transcript variants of the same gene or overlaps between genes. The same genomic region is often assigned to multiple peaks and multiple genes, confounding downstream analyses. Essentially, the program maps ambiguous regions to all genes and all of their transcript variants. For example, exomePeak reports peaks for the gene SLC9A3, as discussed earlier in 3.3.2 Fragment Shifting and Extension, a gene that is otherwise not expressed, because of its overlap with a different gene, BC013821. While only a few cases similar to gene SLC9A3 can be found, the most striking example of finding peaks mapping to a gene that is not expressed, there are likely more examples of genes that have incorrectly adjusted count-data as an artifact from fragmenting shifting. Working within the transcriptomic space has its advantages, but it often hides the problems it attempts to solve. By not making assumptions about peak assignment to genes and transcripts, these edge cases can be solved, or perhaps excluded, during downstream analysis.

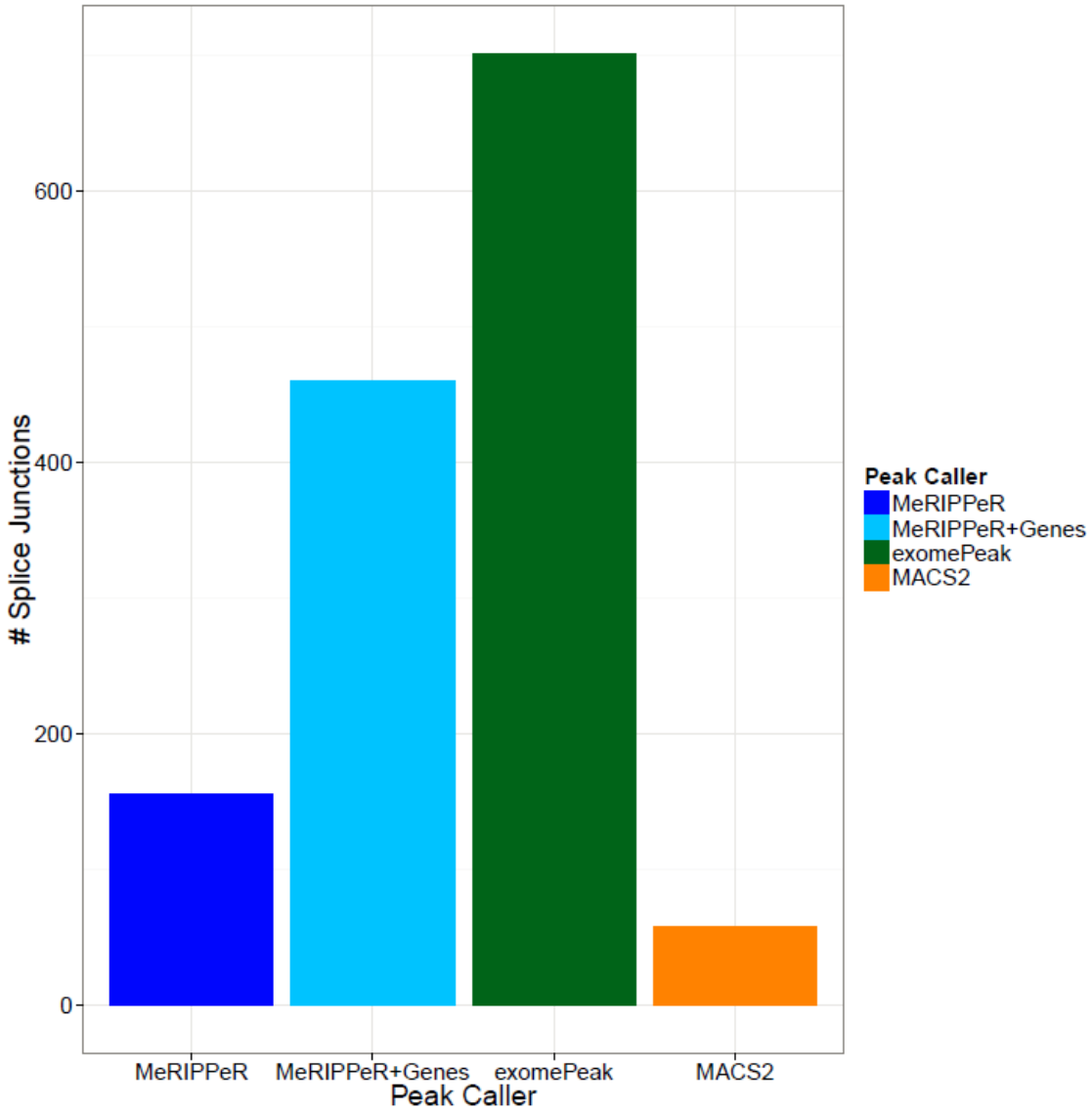


Figure 3.27 Recovery of Splice Junction Peaks Using Spliced Window Augmentation

The number of potentially spliced-peaks, determined by computing how many splice-junctions were spanned by peak regions, shown on y-axis as a function of the peak caller on x-axis. MeRIPPeR+Genes calls a higher number of splice junctions in peaks by using a gene-annotation to augment genomic-based windows with spliced windows.

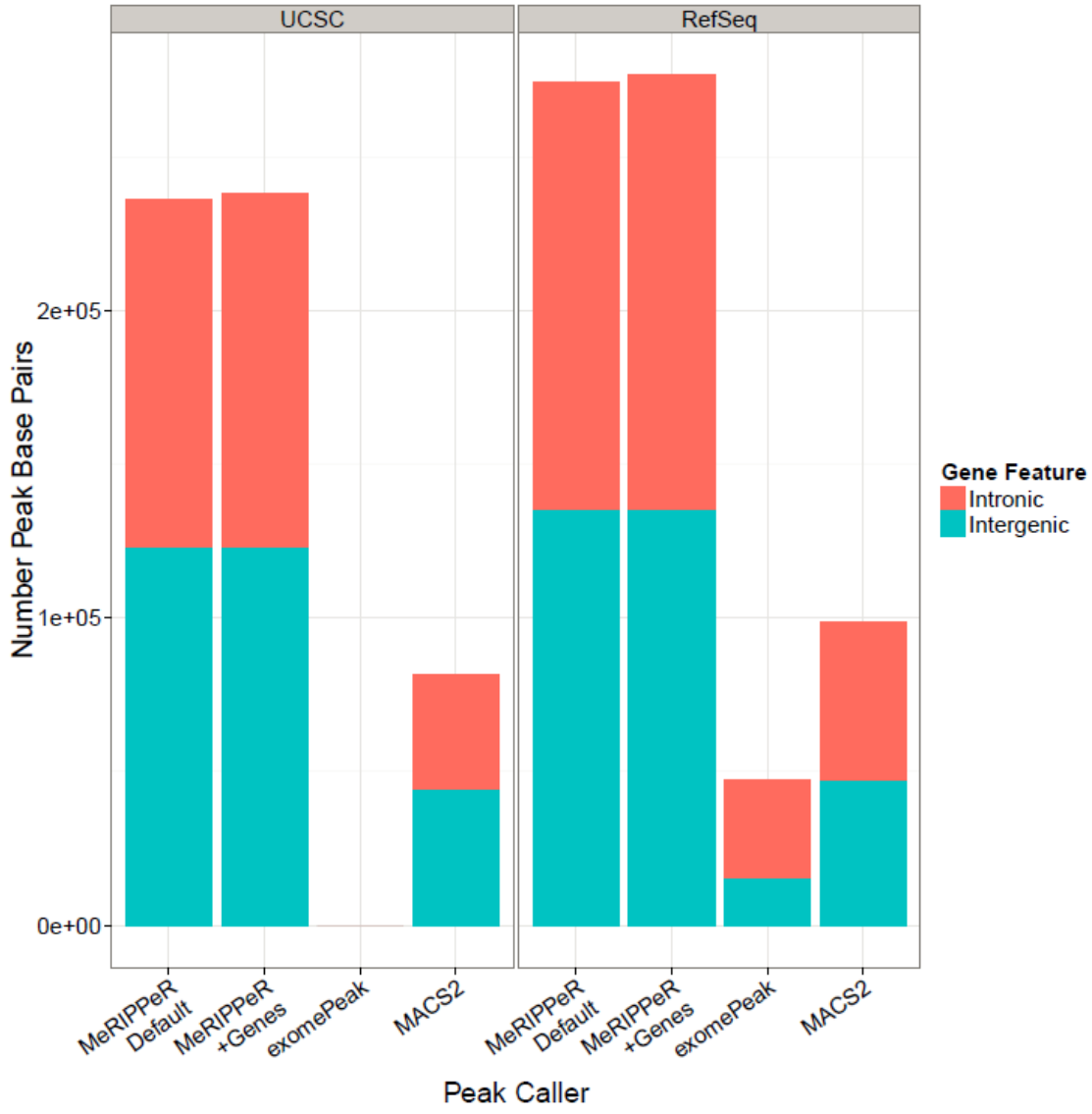


Figure 3.28 exomePeak Captures Little or No Intronic and Intergenic Peaks
The number of peak bases found in intronic (salmon) and intergenic (teal) regions for each peak caller (x-axis), with choice of annotation varying horizontally. The advantage of using MeRIPPeR is that it calls the most peaks in intronic and intergenic regions. exomePeak's regions are fixed to the inputted gene annotation (UCSC) and the results of using RefSeq to annotate these peaks are also shown. MACS2 calls the fewest intronic regions.

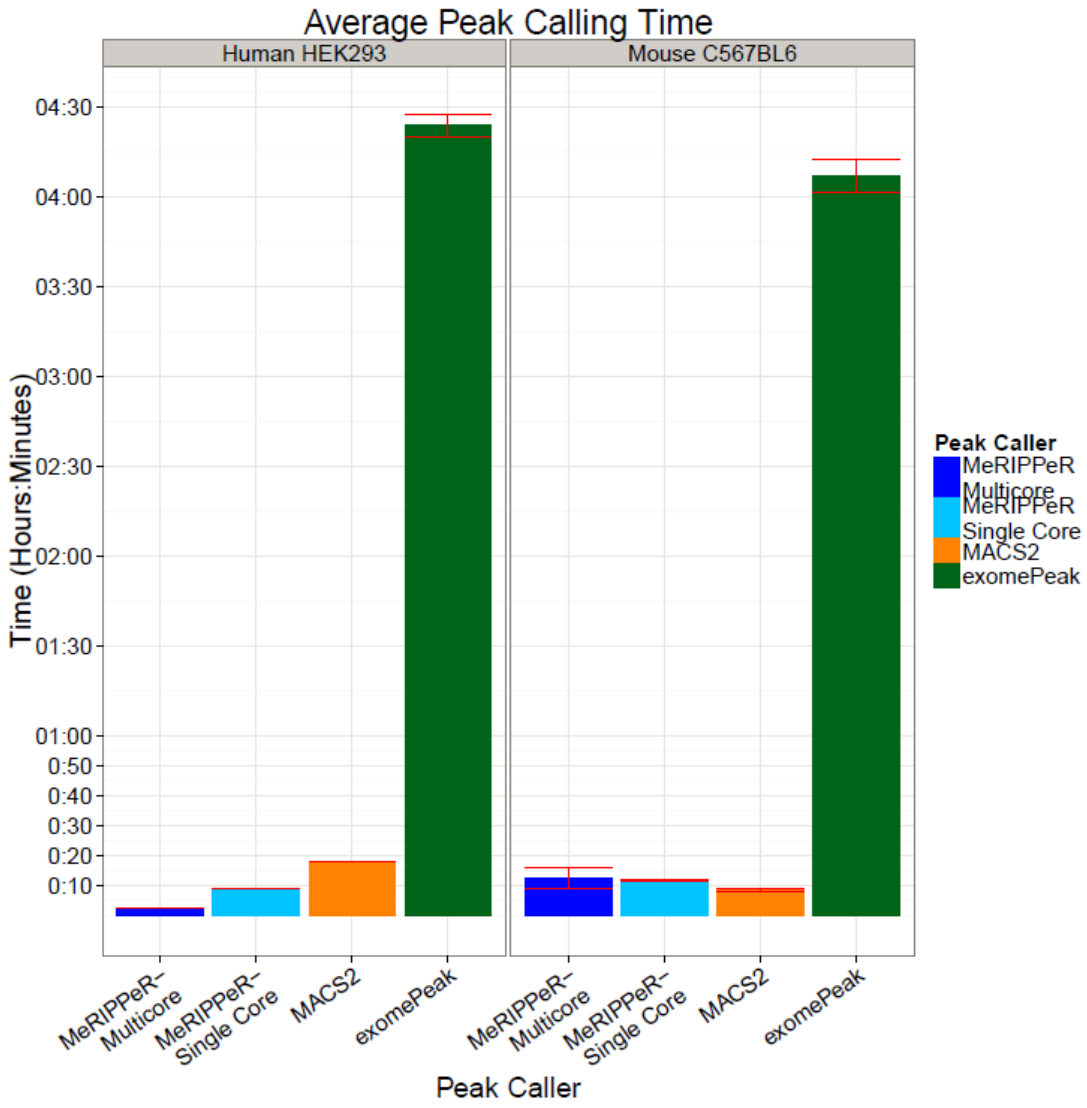


Figure 3.29 MeRIPPeR is Fastest Peak Caller, exomePeak Slowest
Runtime of each peak caller, averaged across five tests, shown on y-axis, as a function of the peak caller, x-axis. MeRIPPeR is the fastest peak caller, built on the *htsjdk* Java package. exomePeak unfortunately performs its coverage calculations in R, which results in it taking nearly 4.5 hours to call peaks, longer than most aligners took to align the same data to the genome.

The *RRACH* consensus motif was not only initially reported to be associated with the METTL3 methyltransferase responsible for methylating adenosine to m^6A, (Harper et al., 1990; Wei and Moss, 1977a) but was also found in peak regions following MeRIP-seq transcriptome-wide mapping. (Dominissini et al.,

2012; Meyer et al., 2012) With such a high degree of specificity of the motif with the methyltransferase, and the presence of the motif in over 80% of the peaks, the motif could also be used to determine the potential true positive rate of each of the peak finders. Moreover, the motif could be used to determine the veracity of the regions in Figure 3.25, showing which peak finders find the most motif-enriched peaks. Figure 3.30 shows the distribution of the number of times the motif appears in the peak regions from Figure 3.25 normalized by peak length, demonstrating that exomePeak has poorer performance than MeRIPPeR and MACS2.

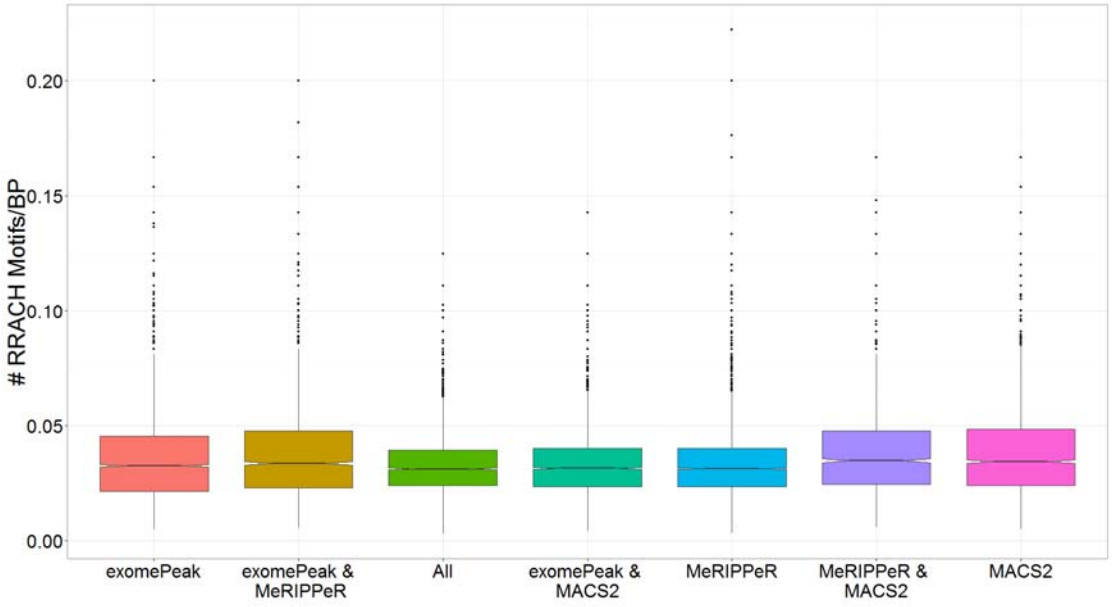


Figure 3.30 exomePeak Performs Worst in Motif Performance
Counting the number of *RRACH* motifs in each of the peak regions in Figure 3.25 and normalizing by each peak length shows exomePeak performs the worst in motif performance, while MACS2 and MeRIPPeR perform comparably.

3.6 Conclusions

Peak calling in MeRIP-seq is highly dependent on a number of variables, including the method of RNA purification, the efficiency of the IP, the number of rounds of IP, and the accurate alignment of the data to the transcriptome. MeRIPPeR is a robust, fast, and powerful peak caller specifically designed for MeRIP-Seq data, but has the potential to be applied to other fragmented RNA-sequencing IP data. Utilizing genomic-based windows, it is not restricted to an annotation set or exonic regions, but augmenting spliced-windows from an annotation database can aid in identifying spliced-peaks. Without ground truth to compare to existing peak callers, using the motif as a measure of truth shows that MeRIPPeR and MACS perform equally well. With augmented annotation support, MeRIPPeR is specifically designed for MeRIP-seq analysis, with higher sensitivity than exomePeak and better sensitivity in spliced peaks than MACS2.

CHAPTER 4 DIFFERENTIALLY METHYLATED PEAK REGIONS (DMPRS)

4.1 Introduction

As was previously mentioned, RNA modifications are dynamic and in addition to varying between tissue and cell types, they also respond to changes in cell stimuli. Global levels of m^6A can be observed to dramatically differ between different tissues (Meyer et al., 2012) and in response to cellular stimuli (Dominissini et al., 2012). Although all the functions of m^6A remain unknown, it has the potential to function as a dynamic layer of translation regulation. Identifying peak regions is the first step in understanding the function of m^6A: determining where in the genome the sites may lie. Its dynamic nature can be further used to elucidate these potential physiological roles. However, these are much harder to quantify than changes in DNA methylation and each challenge must be appropriately addressed.

RRBS and other bisulfide chemistry based assays typically achieve nearly 99% conversion rates. (Garrett-Bakelman et al., 2015) This not only enables a far higher sensitivity in detecting ^5mC sites at single nucleotide resolution over an antibody immunoprecipitation method, but the exact methylation frequency at each base can be estimated as the fraction of cytosines to the total number of cytosines and thymines sequenced at each base. Even with biological and technical variation between samples and replicates, this fraction can be used to accurately find differentially methylated cytosines (DMCs) using logistic regression models. (Akalin et al., 2012)

4.2 Challenges in Identifying Differentially Methylated Peak Regions

However, in the case of m^6A and other RNA modifications, calculating differentially methylated peak regions is far more complicated due to biological and technical variation. The underlying RNA-seq data is observed to have increased biological variance and is often modeled with the negative binomial over the Poisson model to account for this dispersion. (Robinson et al., 2010) This high variance is present in not only the control RNA-Seq samples, but also the MeRIP-Seq, complicating estimation of mean methylation levels.

The total amount of m^6A present in a sample can be calculated using an immunoblot with the same antibody used to perform the IP pulldown. As has been previously demonstrated, the global levels of m^6A are vastly different between different tissues, such as between the brain and the kidneys. (Meyer et al., 2012) Figure 4.1 shows further variation in m^6A levels in mice bone marrow samples, from a collaboration with Ross Levine, MD. The knockout of the Tet2 gene does not have as profound of an impact on global m^6A levels as the VTF samples, which is a knockout model of Tet2 and Flt3. The plate results show biological variance in global m^6A levels between different mice samples, as well as global changes between different types of mice.

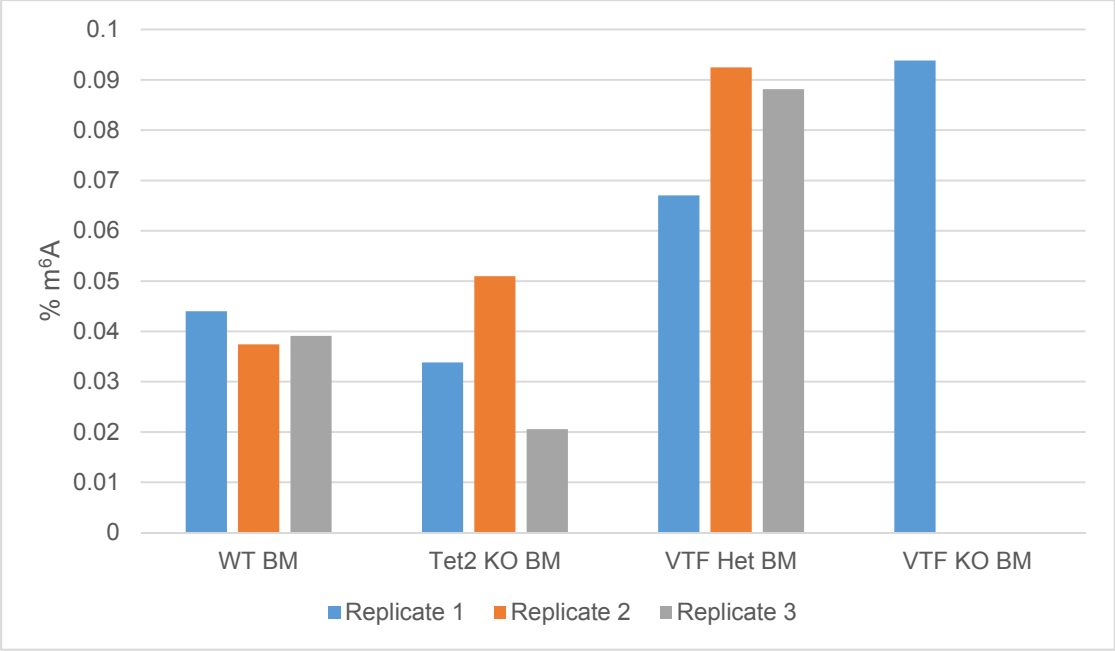


Figure 4.1: Global m^6A Levels in Mouse Bone Marrow Samples Shows Variation Between Sample Types
Percentage of m^6A, normalized by total amount of RNA, in mouse bone marrow (BM) samples obtained from Ross Levine, MD and Alan Shih, MD PhD. Tet2 knockout (KO) had little impact on global m^6A levels relative to VTF heterozygous (Het) and knockout. VTF is both Tet2 and Flt3. Global m^6A levels were measured using EpiQuik m6A RNA Methylation Quantification Kit (Colorimetric) (Epigentek #P-9005-96).

Unfortunately, this global amount of methylation is typically not measured or noted when performing the MeRIP-Seq protocol, but its impact on both peak calling and differential peak calling must be considered. During standard RNA sequencing library preparation, libraries are normalized to equimolar concentrations, to achieve equal coverage across all RNA transcripts. Comparing RNA sequencing data between nuclear and cytosolic fractionated RNA samples, for example, does not reflect that the cytosolic fraction contains more than 2-3 times more RNA. The increase in m^6A content can be the result of either an increase in the amount of m^6A at each site or an increase in the number of m^6A sites. Increased methylation at a single site would result in

higher enrichment observed at that site. However, an increase in the total number of m^6A sites would, in theory, mean more RNA fragments being pulled down in the IP and result in lower global coverage and enrichment. (Robinson and Oshlack, 2010)

In addition, RNA transcript levels themselves are subject to change between replicates and sample conditions. Changes in site-specific m^6A levels must be normalized for these changes. Moreover, differentially methylated regions should have significant changes in the fraction of RNA that is methylated at each site. This fraction is more challenging to estimate, as it is dependent on the accurate estimation of both the RNA-Seq and MeRIP-Seq levels in each window. The MeRIP-Seq fraction is highly variable based on the reproducibility of the IP, as well as biological variation in m^6A sites.

Technical variation in MeRIP-Seq can affect both peak calling and differential peak estimation. The efficiency of the IP is dependent on numerous factors, including the IP binding conditions and the specificity of the antibody. Performing all of the IPs in a single batch can reduce some technical variation, but as discussed earlier in 3.4.2 MeRIP-Seq IP Enrichment, some technical variation will continue to exist between two different replicates.

Lastly, m^6A is a single-nucleotide modification while the resolution of the IP is closer to 100-200 base pairs. Each peak has the potential to encompass many methylation sites, especially when some peaks can span up to 1,000 base pairs or more. (Meyer et al., 2012) This lack of resolution complicates determining site-specific methylation changes that might occur. A significant increase or decrease in methylation at a single site could be obscured by lack of changes in methylation in sites surrounding it. Without explicit single-nucleotide

resolution, such as in DNA methylation, the statistical power to find DMRs is significantly reduced.

4.3 Existing Methods in Detecting Differentially Methylated Peak Regions

Prior to the development of methods to achieve transcriptome-wide mapping of m^6A, m^6A was already known to be a dynamic modification subject to significant changes in response to treatments. (Clancy et al., 2002) The first method to attempt to look at differentially methylated peak regions in m^6A was performed by Dominissini et al. when they explored the effect of multiple treatments on m^6A sites in the HepG2 cell line. They determined that 70-95% of methylation sites remained largely unchanged between treatments but were able to find a small subset of peaks that did change between treatments, though they admitted that their methods were likely very conservative. (Dominissini et al., 2012)

Following the publication of the exomePeak Bioconductor package, additional methods were built into the method to calculate differentially methylated peak regions. The exomePeak method of calling peaks uses a Poisson distribution to model the read counts in each of the samples. (Meng et al., 2013) They scale the read counts to normalize for the number of reads mapped, show that the data follows the hypergeometric distribution, and use Fisher's Exact Test to test for significance, denoted *RHtest*. (Meng et al., 2014) Two additional methods were then developed to detect differential methylation within the exomePeak package. *RHHMM* uses a Bernoulli Hidden Markov-Model (HMM, which they claim improves spatial resolution. (Zhang et al., 2014b) The binomial likelihood ratio test, *bltest*, models the data using a binomial distribution and finds differentially methylated regions by comparing the means. (Zhang et al., 2014a)

Unfortunately, the methods are very conservative, with both the *RHtest* and *bltest* reporting the same five differentially methylated regions in the FTO knockout mouse study. (Hess et al., 2013) The methods are also very slow, taking over six hours to find the differentially methylated sites.

4.4 Methods

Unfortunately, most of the challenges mentioned earlier in identifying differentially methylated peak regions are difficult to solve. Normalizing for IP efficiency, for example, is confounded by biological variation within the m^6A sites. Technical replicates, in addition to biological replicates, would be preferred, but the costs and input limits are too great to implement them in an experimental design. Nevertheless, the best approach is to attempt to normalize for some of them, and use quality control metrics to exclude replicates with technical artifacts.

Furthermore, the resolution of MeRIP-seq is on the order of hundreds of base pairs, far lower than the single-nucleotide modification. A peak region larger than 200 base pairs likely spans multiple m^6A sites, but without sufficient data, the peak cannot be accurately split into regions encompassing a single site. This complicates differential methylation analysis, in that only part of a peak, representing a single or perhaps small cluster of m^6A sites, may change in response to external stimuli. Examining the data at full-peak resolution might mask smaller minute changes. To work around this problem, the windowed approach, used earlier in Chapter 3.4.2 MeRIP-Seq IP Enrichment to demonstrate window enrichment, uses multiple overlapping 100 base pair windows that step at 25 base pairs within peak regions to detect smaller changes in large peaks.

EdgeR, limma, DESeq and other Bioconductor packages have been used multiple times with great success in identifying differentially expressed genes (DEGs). (Anders and Huber, 2010; Ritchie et al., 2015; Robinson et al., 2010) While they theoretically may be used to identify peak regions genome-wide, their implementations largely limit their use to annotated genomic regions, on the order tens of thousands of genes, for example. Windows within peak regions, on the other hand, are far smaller in scale and could be used effectively in conjunction with these tools to identify differentially methylated peak regions. The tools are already designed to handle RNA-seq data, including normalizing for sequencing depth, number of genes expressed, and fitting a linear model. The methods below will discuss using EdgeR, but the same methods could be applied using *limma* or *DESeq*. The edgeR Bioconductor package most notably models the underlying RNA-sequencing data using the negative binomial to model the dispersion caused by biological variation in the data. This dispersion, in particular, is useful in modeling the increased variance caused by the IP.

Applying the method with the FTO mouse knockout data (Hess et al., 2013), there are two types of mice, wild type and FTO knockout mice, with three biological replicates of each. MeRIP-seq was performed on each of them, producing six IP samples and six control samples, for a total of 12 sequencing samples. The original library sizes were used, to prevent scaling RNA-seq libraries to MeRIP-seq depth, and TMM scaling was applied, though separately to the IP and Control samples, and merged later, to prevent the same thing occurring there. The distribution of the log 2 enrichment without applying TMM scaling is shown in Figure 4.2, which shows a high degree of variation in the IP. Applying TMM scaling separately to the IP and control samples, the log 2 distribution of adjusted enrichment counts appears more consistent, depicted in

Figure 4.3. The data was modeled within edgeR with the IP nested within the mouse type, and differential methylation was applied as a contrast matrix comparing the two IP samples. A volcano plot of the results are shown in Figure 4.4, which compares the log 2 fold change relative to the –log 10 of the raw p-value. Only four sites actually pass the significance cutoff after p-value adjustment, though the windows come from the same peak region. They are colored in red on Figure 4.4 and summarized in more detail in Table 4.1. The region corresponds to a predicted Ensembl gene ENSMUST00000083437.

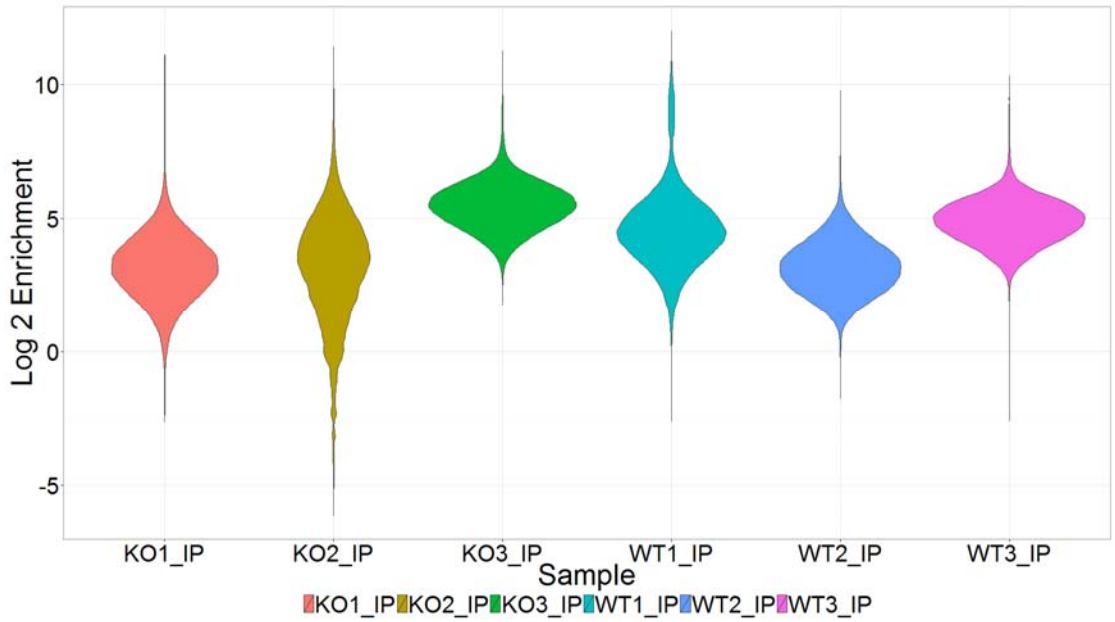


Figure 4.2 Distribution of Log 2 Enrichment without Scaling Shows Technical Variance in IP Efficiency
Without applying TMM scaling to scale the control samples and, especially, the IP samples, the log 2 enrichment distributions show strong differences in IP enrichment. FTO Knockout samples are shown as KO, wild type samples as WT, and replicate denoted by number.

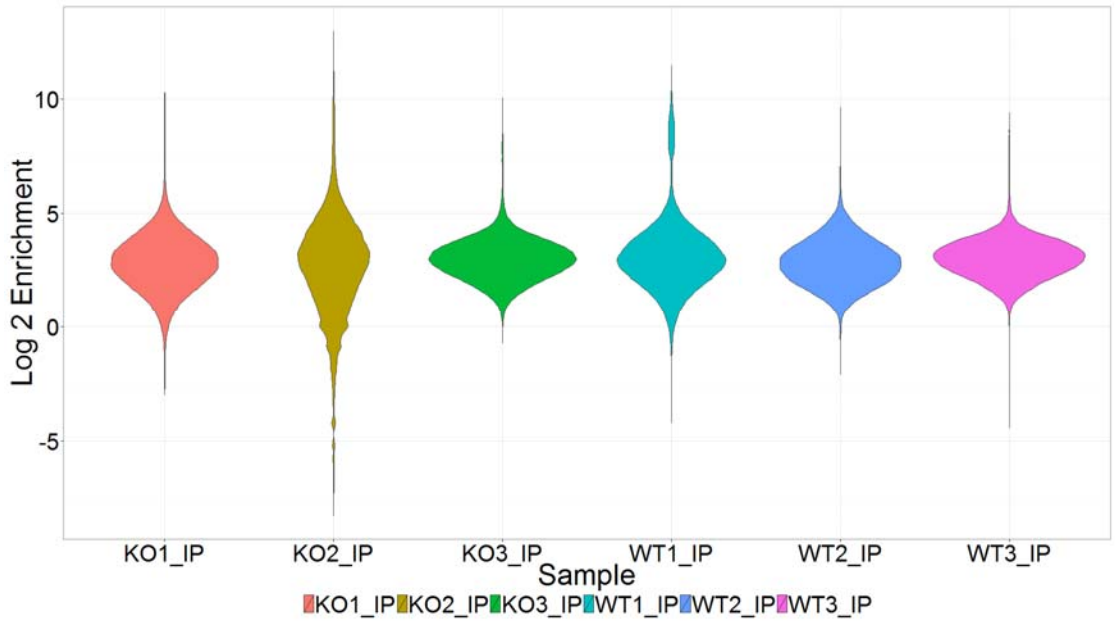


Figure 4.3 TMM Scaling Normalizes for Technical Variance in Enrichment
Scaling the read counts for the IP and control samples separately normalizes the samples for differences in coverage and IP efficiency. FTO Knockout samples are shown as KO, wild type as WT, and replicate denoted by number.

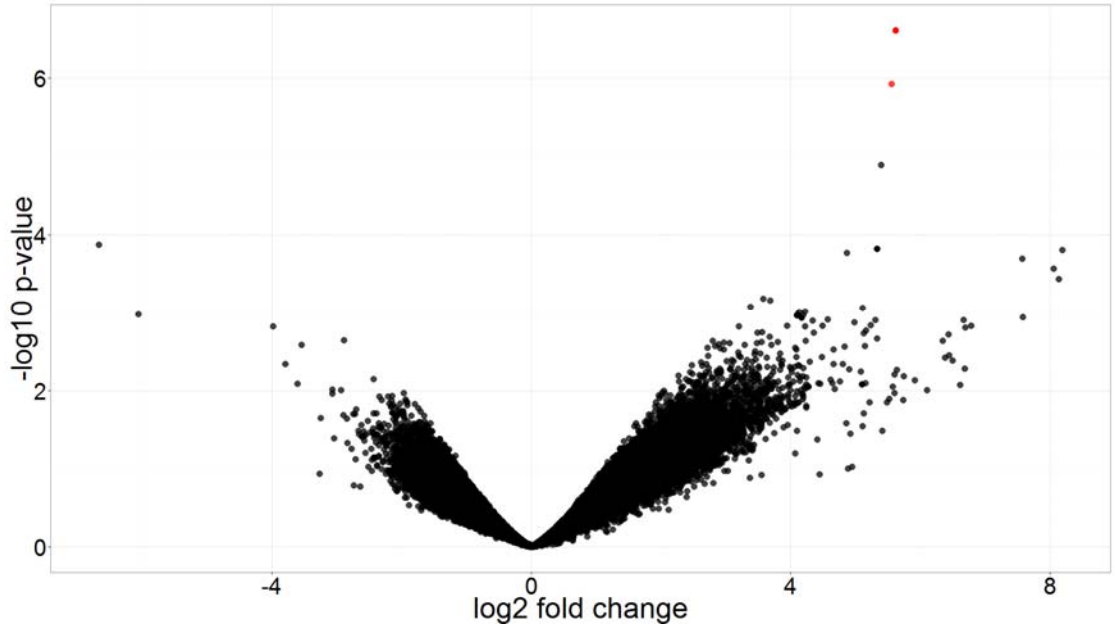


Figure 4.4 EdgeR Differential Methylation Analysis Captures Two DMPRs
Volcano plot showing the log $_2$ fold change versus the –log 10 p-value of the methylation difference in FTO/WT. The raw p-value is plotted; only two peak regions are significant after Benjamini-Hochberg adjustment, denoted in red.

Table 4.1: Differentially Methylated Peak Regions in FTO KO Data (edgeR)

CHR	START	END	LOGFC	P-VALUE	FDR
chr10	4484275	4484325	5.618325	2.45E-07	0.008541
chr10	4484250	4484325	5.618325	2.45E-07	0.008541
chr10	4484225	4484325	5.618325	2.45E-07	0.008541
chr10	4484300	4484325	5.558247	1.18E-06	0.030866

4.5 Conclusions

Although the method developed using edgeR failed to capture significantly more differentially methylated regions than pre-existing methods, the TMM scaling applied to the distribution of log 2 enrichments showed that the method can be used to effectively account for differences in IP efficiency. In addition, the p-value adjustment method uses the Benjamini Hochberg method, which assumes statistical independence between the tests. Since the windows are overlapping at 25 base pairs, this assumption of independence is no longer valid, and the adjustment method is likely over-correcting the family-wise error rate. Using a lower cutoff could yield better results but a higher false positive rate. Existing methods, such as the *bltest* and *rhtest*, capture very few sites, as well, but do not account for technical variance in the IP.

CHAPTER 5 THE FUNCTIONAL AND PHYSIOLOGICAL ROLE OF METHYL-6-ADENOSINE IN RESPONSE TO HEAT SHOCK AND RIBAVIRIN

5.1 Introduction

Without knowing the function of all of the writers, erasers, and readers of m^6A, its full physiological role remains unknown. YTDF2 was recognized as one of the potential readers, and determined to mediate mRNA decay through p-bodies. (Wang et al., 2014a) Yet m^6A sites are heavily found in nuclear RNA, (Levis and Penman, 1978) with its demethylase FTO being found in nuclear speckles. (Jia et al., 2011) This implies m^6A has the potential for multiple roles in gene expression and regulation, further implicating it in splicing (Dominissini et al., 2012; Saletore et al., 2013; Zhao et al., 2014) and perhaps even nuclear export. Further examining m^6A sites in nuclear fractionated RNA would serve to answer these questions.

The antiviral drug ribavirin is a guanosine analogue and traditionally used to terminate viral RNA synthesis. (Kentsis et al., 2004) Viral RNA-dependent RNA polymerases incorporate ribavirin in place of guanosine, leading to viral mutagenesis. (Crotty et al., 2002) In addition, ribavirin was shown to bind to and inhibit the eukaryotic translation initiation factor eIF4E, (Kentsis et al., 2004; Kentsis et al., 2005) which itself is responsible for recruiting mature mRNA transcripts to ribosomes via the 5' 7-methyl-guanosine cap. (Gingras et al., 1999) eIF4E is often found elevated in cancer cells, and ribavirin's inhibition of its activities was further shown to aid in the treatment of acute myeloid leukemia. (Assouline et al., 2015; Assouline et al., 2009; Borden and Culjkovic-Kraljacic, 2010) Its repression of the nuclear export of specific mRNAs enables it to be specifically used in a controlled experimental design to determine its

effect on m^{6}A sites and their potential relationship with the genes that are repressed.

Furthermore, heat shock of cells has been well-studied and characterized as a method of cellular stress, including its activation of the heat shock proteins. (Lindquist and Craig, 1988) Its effect on m^{6}A sites was initially examined among other treatments, (Dominissini et al., 2012) and through immunoblots was found to a dramatic increase on global levels of m^{6}A.[1] The immediate export and translation of the heat shock proteins could be further used to examine the role of m^{6}A in nuclear RNA.

5.2 Methods

The full experimental design served to compare three groups, a control group of untreated cells, cells treated with heat shock, and cells treated with ribavirin. Each group would consist of total RNA samples, as well as RNA from nuclear and cytosolic fractions. Three replicates of each would be used, and MeRIP-Seq would be performed on each of the samples. In order to meet the minimum input requirements for a single round of IP, 50 micrograms of RNA was required at a minimum, requiring hundreds of millions of cells to be harvested specifically for the nuclear fractionated RNA. Cells from the Ly1 diffuse large B-cell lymphoma (DLBCL) cell line were collected by Tharu Fernando, MS of the Ari Melnick, MD laboratory, working in collaboration with Leandro Cerchietti, MD. Over 1.4 billion cells were required in total to meet the nuclear fractionated RNA requirement, and split into 3 fractions for the control, ribavirin, and heat shock treatment, before being separated into their individual RNA fractions. The RNA samples were fractionated by Tharu Fernando using standard nuclear RNA

[1] Unpublished data from Kate Meyer, PhD with the Samie Jaffrey, MD PhD laboratory, in collaboration with the Ari Melnick, MD laboratory.

extraction protocol. Heat shocked samples were heated to 43° C for two hours and allowed to recover at 37° C for two hours. Ribavirin cells were treated with 100 μM ribavirin for four hours. The full experimental design is depicted in Figure 5.1.

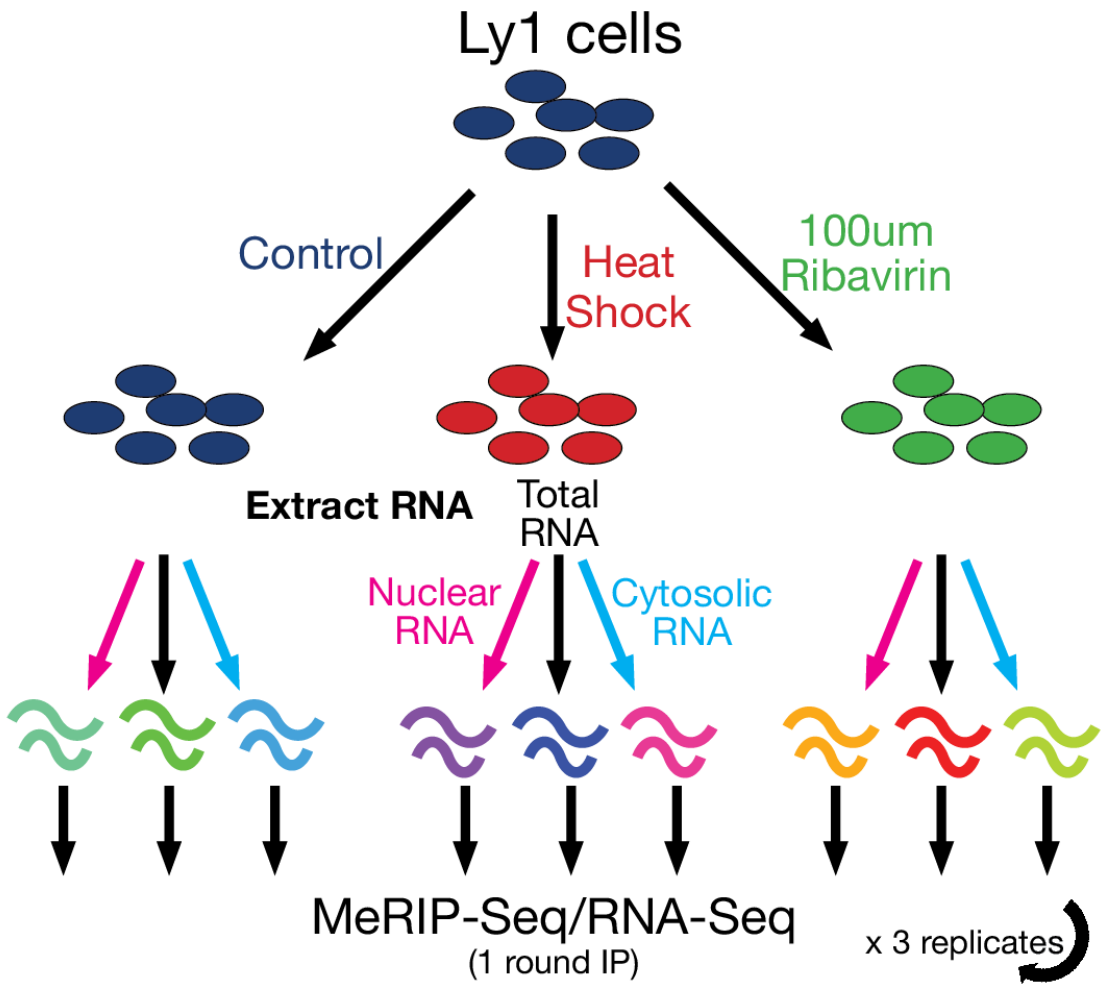


Figure 5.1 Heat Shock and Ribavirin and Nuclear vs Cytosolic MeRIP-Seq Experimental Design.
A full 3x3x3 design with MeRIP-seq and RNA-seq samples resulting in a total of 54 samples. Starting with a total of 1.25 billion cells split into three fractions, one control, one treated with heat shock, and one treated with ribavirin. Each would be further split into three fractions of RNA, total, nuclear, and cytosolic, each itself in replicates of three. MeRIP-seq and the control RNA-seq libraries were then prepared for each replicate for a total of 54 samples.

With 50 micrograms of input material, only a single round of IP was performed to ensure all samples resulted in successful libraries, using polyA-purification to remove rRNA contamination. Unfortunately, one of the nuclear fractionated heat shock replicates was not successful, but its control RNA-seq could still be used to estimate RNA-seq levels. The cDNA libraries were then sequenced on an Illumina HiSeq 2500 at single-ended 42 base pairs.[2]

The sequencing data was aligned to the hg19 genome (excluding haplotype and random chromosomes) using the STAR (Dobin et al., 2013) aligner, excluding multimapped reads and ensured all samples met quality control specifications using R-Make. Peaks were called using MeRIPPeR with augmented junction annotation support from RefSeq gene annotations. Peaks were required to be present in at least two out of three of the replicates, as opposed to the default of requiring it to be present in all replicates, to account for one of the samples only having two replicates present.

5.3 RNA-Sequencing Analysis

Results from R-Make (Li et al., 2014a) in Figure 5.2 show that the samples were not pooled well and read mapping counts were not equally distributed, with one of the ribavirin cytosolic MeRIP samples getting fewer reads. The read mapping distribution to gene features was previously discussed in Figure 2.2 as proof that polyA-purification of samples results in lower rRNA contamination when working with higher RNA input samples. More reads can also be observed to map to intronic segments in the nuclear fraction, as can be expected with

[2] The samples were dual-indexed using a high throughput (HT) Illumina sequencing kit. Unfortunately, Illumina does not sell SBS V3 kits with enough bases to sequence the dual indexes, so 8 base pairs of the standard 50 base pairs had to be used to sequence the second index, leaving only 42 base pairs for each read. SBS V4 kits do have enough reagents but are not compatible with the Epigenomics Core sequencers.

nuclear fractionated RNA, showing the presence of pre-spliced immature mRNA transcripts. An MDS plot of the samples in Figure 5.3 depicts clear separation along the first dimension corresponding to the fraction and the second dimension shows separation along the treatments, though the Ribavirin treatments do not show a significant change in RNA-Seq from the control.

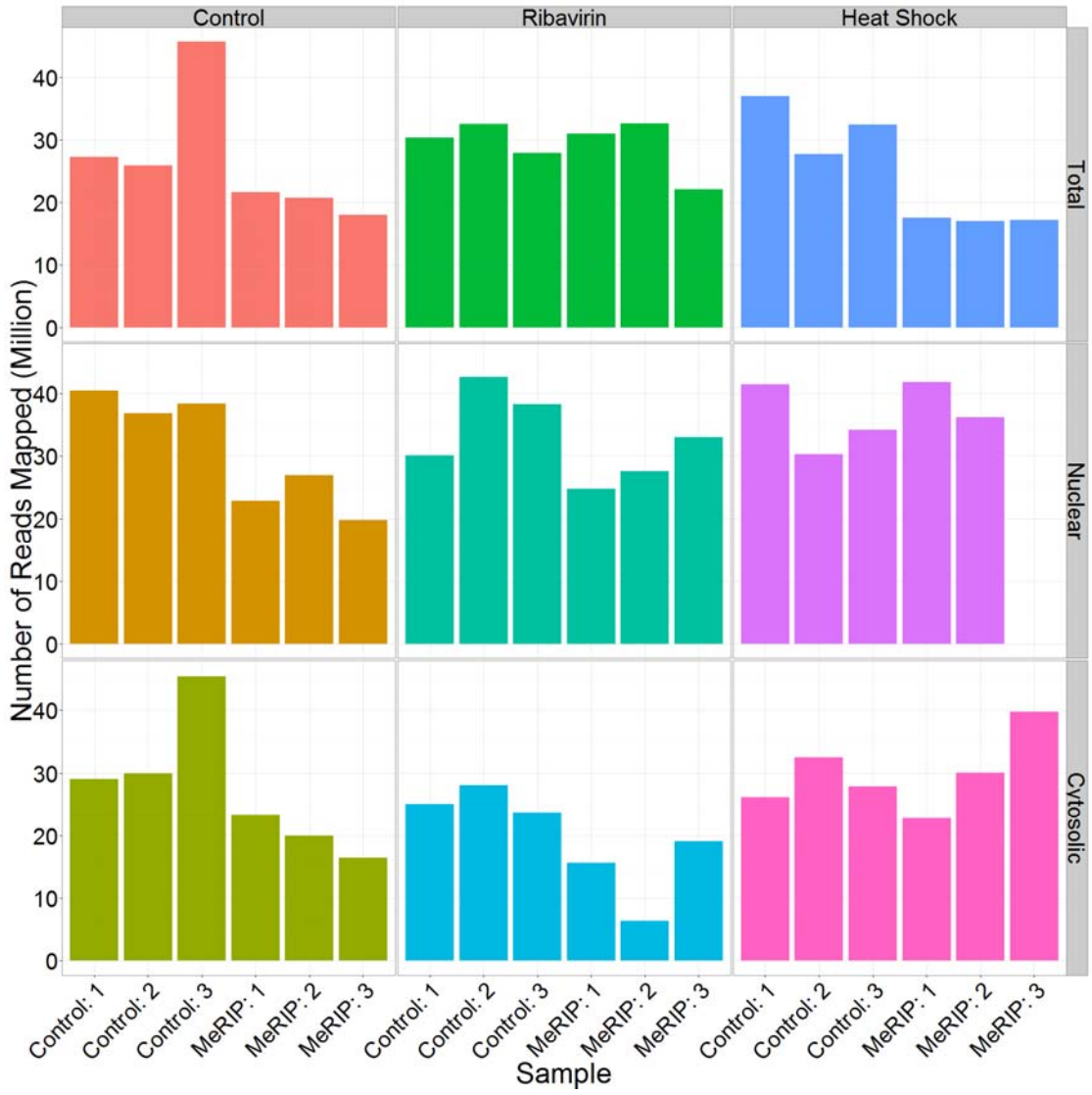


Figure 5.2 Heat Shock and Ribavirin Read Mapping Distribution
Distribution of the number of reads mapped (in millions) to each of the samples. The third nuclear replicate MeRIP sample failed, which is blank.

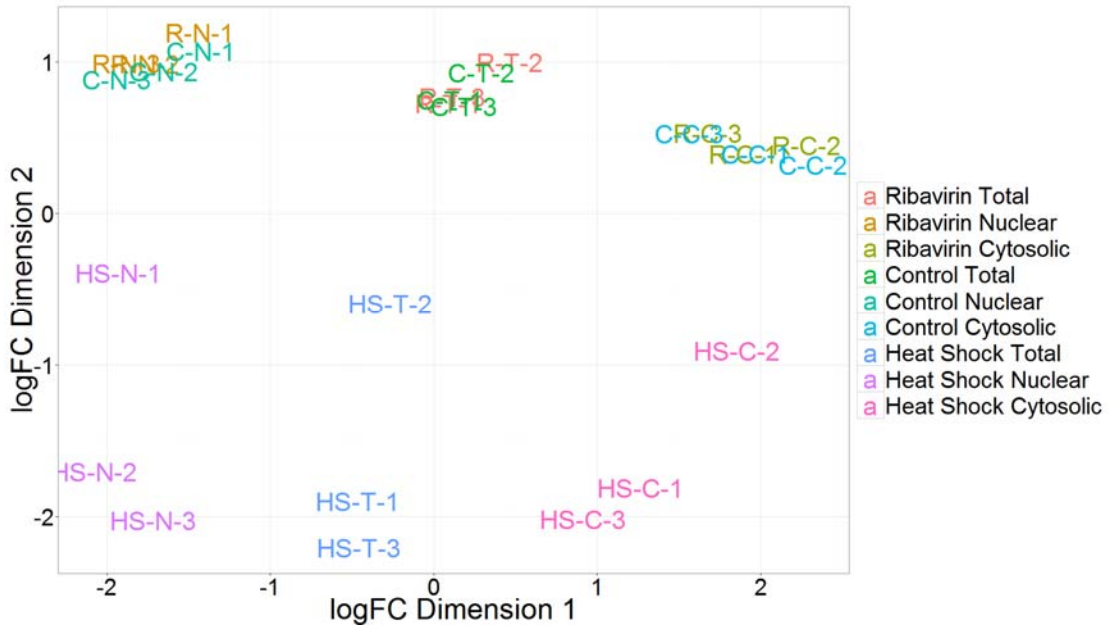


Figure 5.3 MDS Plot of RNA-Seq data Shows Separation of Fraction and Heat Shock
MDS Plot of RNA-Sequencing data shows a strong separation on the first dimension with respect to the fraction. The second dimension shows separation of the heat shock samples, but the Ribavirin treatments did not affect the RNA-Seq levels as much as expected

However, Ribavirin affects the nuclear export of specific genes, which may be masked when examining all of the genes together. In addition, the ideal control for the Ribavirin treatment would have been a vehicular control sample. In the absence of such a sample, the untreated control sample will have to serve as a control. Although this is not ideal, the fact that the samples are globally similar to the Ribavirin treated samples in the MDS plot shows that they can still be used as an adequate control.

5.3.1 Heat Shock RNA-Sequencing Analysis

Unsurprisingly, a volcano plot showing the log 2 fold change versus the -log 10 p-value of the difference between the heat shock and control samples in the total RNA sequencing data in Figure 5.4 shows a large fraction of the genes can

be classified as *differentially expressed*, a total of 1,898 genes out of the 12,508 genes that were determined to be expressed in the samples[3]. A heat map of the log 2 fold change in expression in genes in the HSP70 gene families shows a dramatic increase in HSP70 genes, depicted in Figure 5.5, most of which are differentially expressed in all of the fractions. HSPA12A and genes from the HSP90 gene family were not in the expressed set of genes and were excluded.

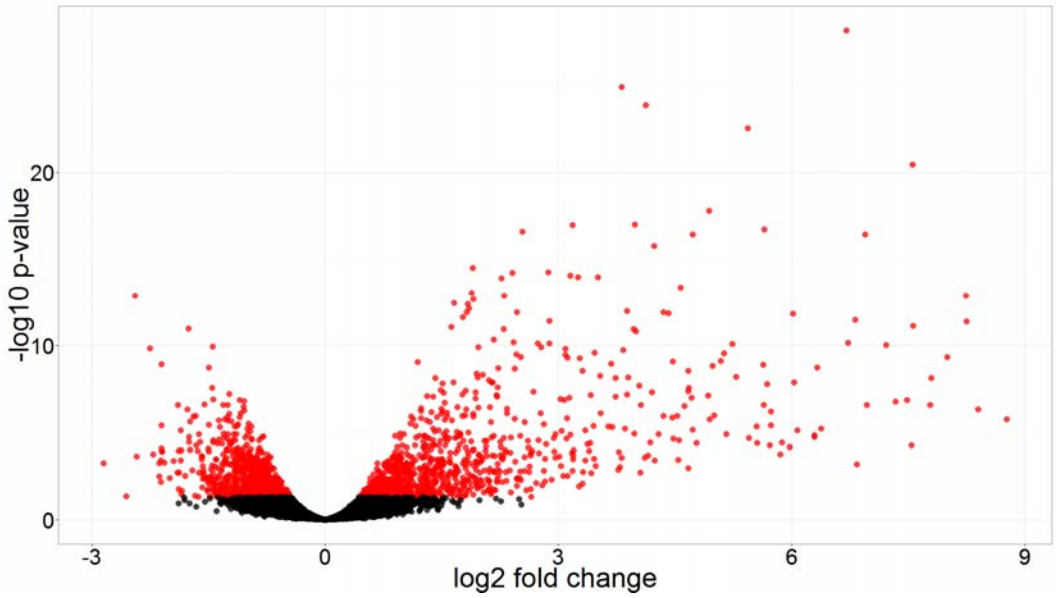


Figure 5.4 Volcano Plot of Heat Shock Total RNA Shows Many Upregulated Genes
Comparison of total RNA-sequencing data between the heat shock and control samples shows a large number of differentially expressed genes, shown in red, many of which are significantly up-regulated in response to the heat shock.

[3] Gene expression was determined by requiring a counts per million (CPM) of at least 1 in at least 6 of the RNA-Sequencing samples.

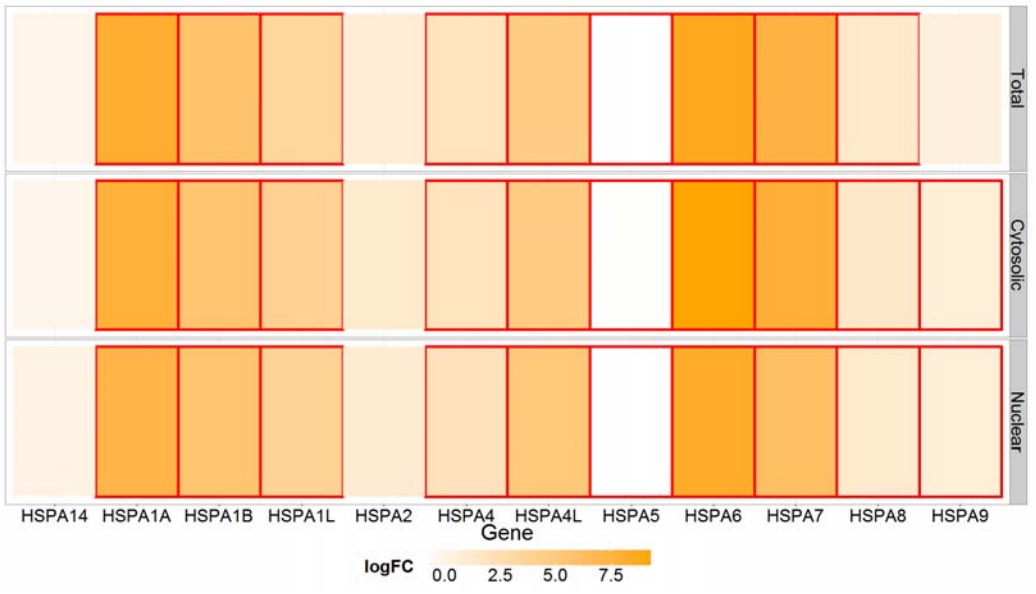


Figure 5.5 Heat Map of Log 2 Fold Change of HSP70 Genes
Heat map of the log fold change in genes of the HSP70 gene family shows a dramatic increase in their expression relative to the control. Differentially expressed genes are denoted in red, log fold change shown for each fraction. The changes are consistent across all fractions.

Figure 5.6 illustrates the concordance of differentially expressed gene sets in between the total RNA and nuclear and cytosolic RNA fractions. Unique subsets of genes can be found to be differentially expressed in each fraction, with 950 genes common to all three of them. Gene ontology analysis of the up-regulated genes shows an enrichment in both stress response and heat shock, shown in Table 5.1 using DAVID. (Huang da et al., 2009a, b)

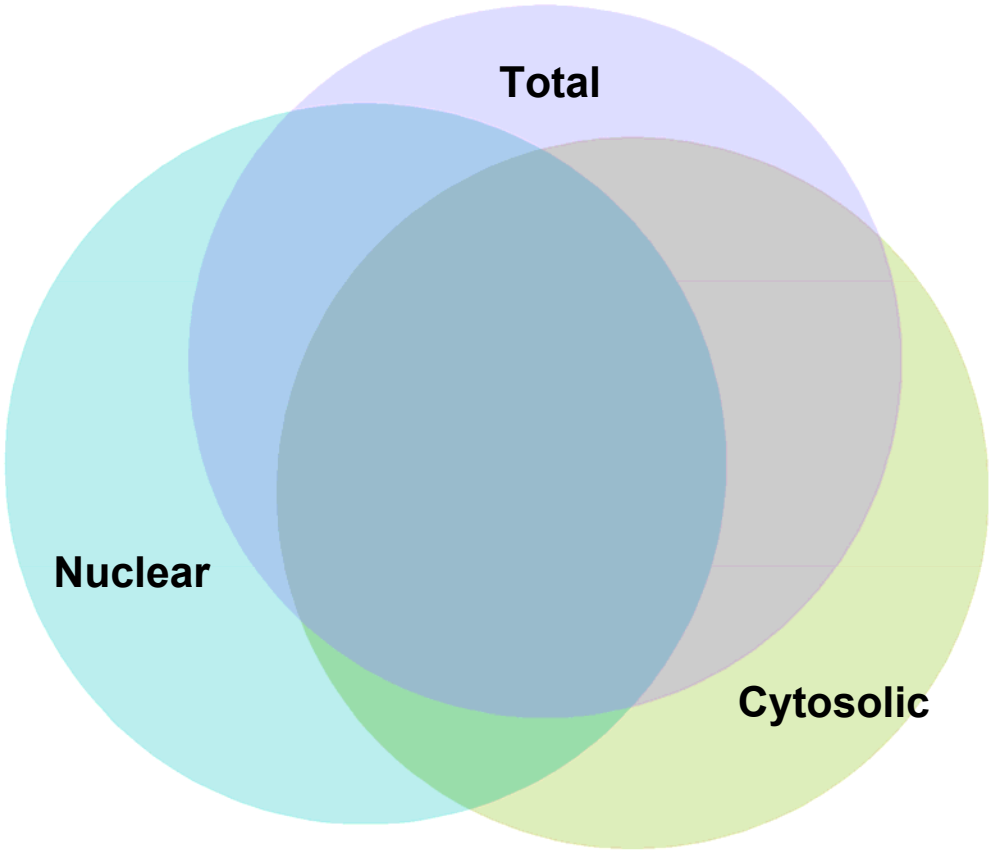


Figure 5.6 Venn Diagram of Differentially Expressed Genes in Fractions Shows High Number of DEGs Common to All Fractions
A Venn diagram showing the overlap of differentially expressed genes found in the total RNA-seq and nuclear and cytosolic fractions. Unique subsets of genes can be found to be differentially expressed in each subset.

Table 5.1: Functional Annotation of Up-Regulated Genes in Heat Shock (Total RNA) using DAVID

Biological Pathway	Benjamini P-Value
phosphoprotein	6.8E-10
stress response	1.2E-09
Cytoplasm	2.2E-08
response to unfolded protein	1.4E-07
protein folding	1.2E-06
molecular chaperone	1.6E-06
Chaperone	1.8E-06
response to protein stimulus	0.00002
unfolded protein binding	0.000042
alternative splicing	0.0005
positive regulation of programmed cell death	0.002
regulation of apoptosis	0.002
positive regulation of cell death	0.002
positive regulation of apoptosis	0.0021
cell death	0.0021
Death	0.0022
regulation of cell death	0.0024
regulation of programmed cell death	0.0025
Apoptosis	0.0039
programmed cell death	0.0054
Dioxygenase	0.008
response to organic substance	0.0087
intracellular signaling cascade	0.0098
protein kinase cascade	0.012
splice variant	0.013
heat shock	0.02
atp-binding	0.02
Nucleus	0.022
protein amino acid phosphorylation	0.037
nucleotide-binding	0.042
mutagenesis site	0.044
compositionally biased region:Glu-rich	0.046
developmental protein	0.05

The fractionated data can be used to identify genes that have significant changes in their nuclear to cytosolic ratios, the volcano plot of which is depicted in Figure 5.4. Although a fair number of genes are deemed *differentially*

exported, pathway analysis does not indicate an enrichment of any specific pathways. The differentially exported genes and their log fold change in the nuclear/cytosolic ratio is shown in Figure 5.8. However, the analysis of differential export is more suited for the Ribavirin treated samples, although it can still be examined in the context of the heat shock treatments.

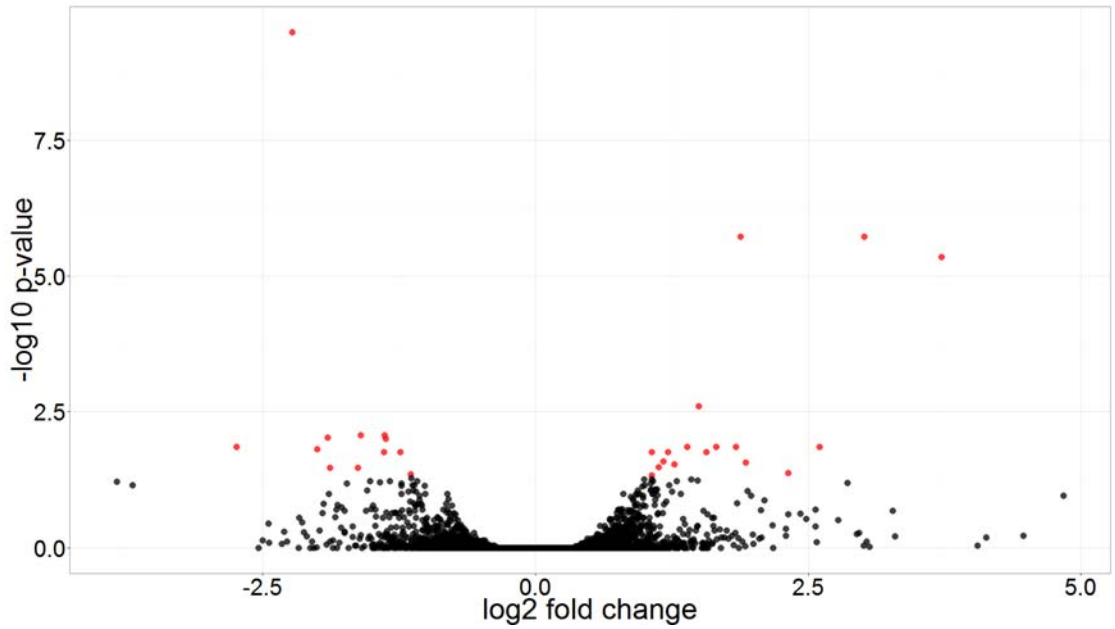


Figure 5.7 Heat Shock Induces Some Significant Changes in Nuclear/Cytosolic Ratio
A volcano plot showing the log 2 fold change in the change in the fraction of nuclear to cytosolic levels and the -log 10 adjusted p-value. Genes showing significant nuclear/cytosolic ratio are shown in red.

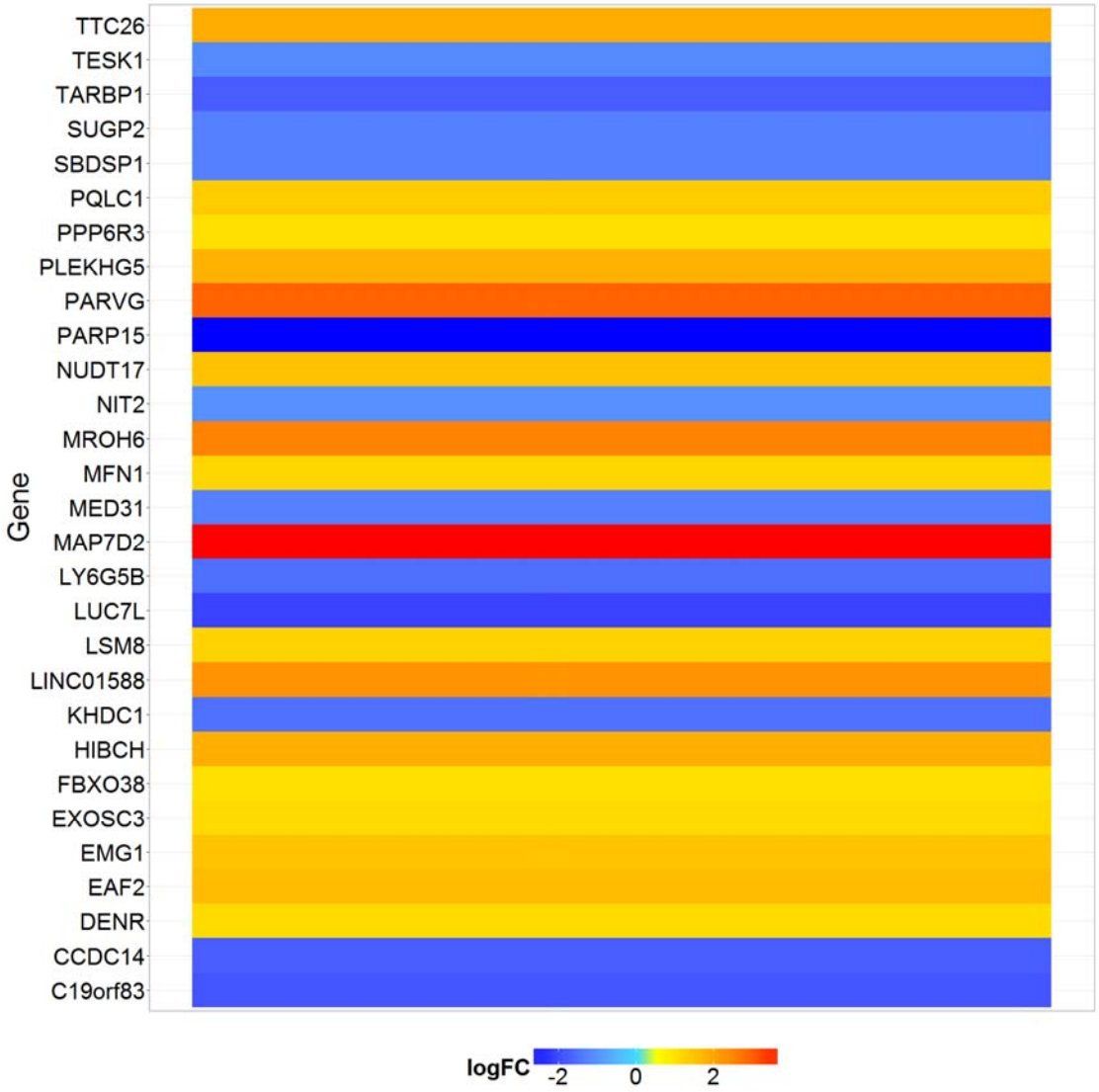


Figure 5.8 Log Fold Change in Heat Shock Nuclear to Cytosolic Ratio
A heat map showing the log fold change of the nuclear to cytosolic ratio for genes that were determined to be differentially exported.

5.3.2 Ribavirin Treatment RNA-Sequencing Analysis

The multi-dimensional scaling plot in Figure 5.3 showed very little change in the RNA-sequencing data in the Ribavirin treated samples, relative to the control. A principal components analysis (PCA) of only the Ribavirin and control samples, excluding the heat shock samples, shown in Figure 5.9, shows the first two

dimensions separate the samples along the fractionation. The Ribavirin treated samples still remain tightly clustered with the control samples. A scree plot of the variance in the PCA analysis, Figure 5.10, shows that these first two dimensions capture the vast majority of the variance, and additional dimensions do not result in clear separation of the Ribavirin treatments. Consequently, no genes are found to be differentially expressed in any of the fractions, so only a smear plot of the average log counts per million of each gene and its log fold change in the Ribavirin treatment is shown in Figure 5.11 for reference.

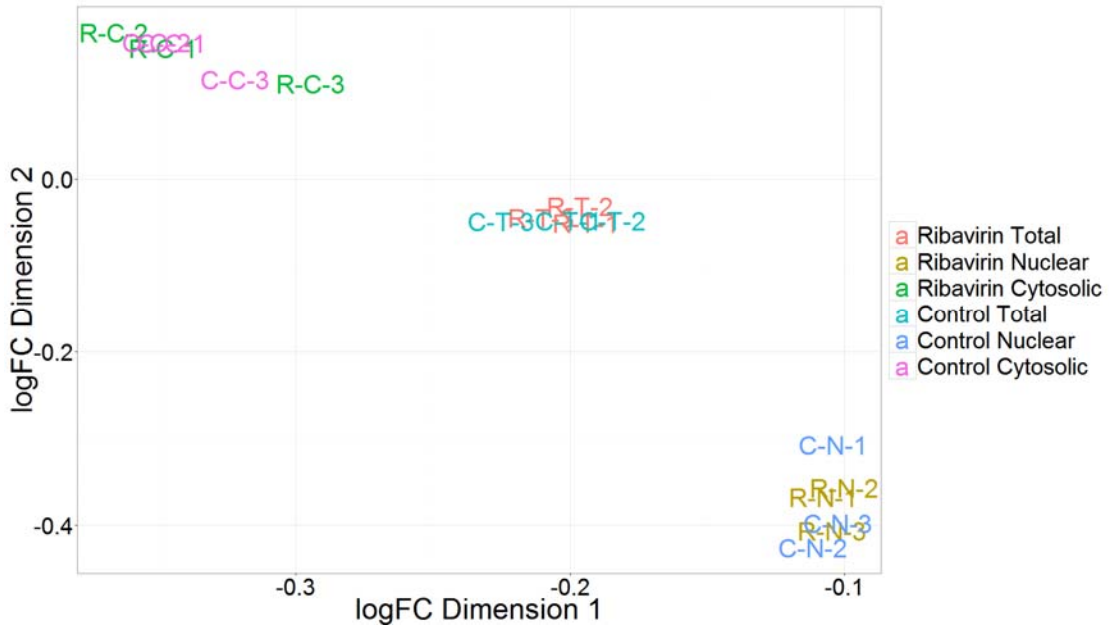


Figure 5.9 Ribavirin and Control Samples Remain Tightly Clustered
Principal Component Analysis of the counts per million in the Ribavirin and control samples separates the samples on the first and second dimensions along the fractionation.

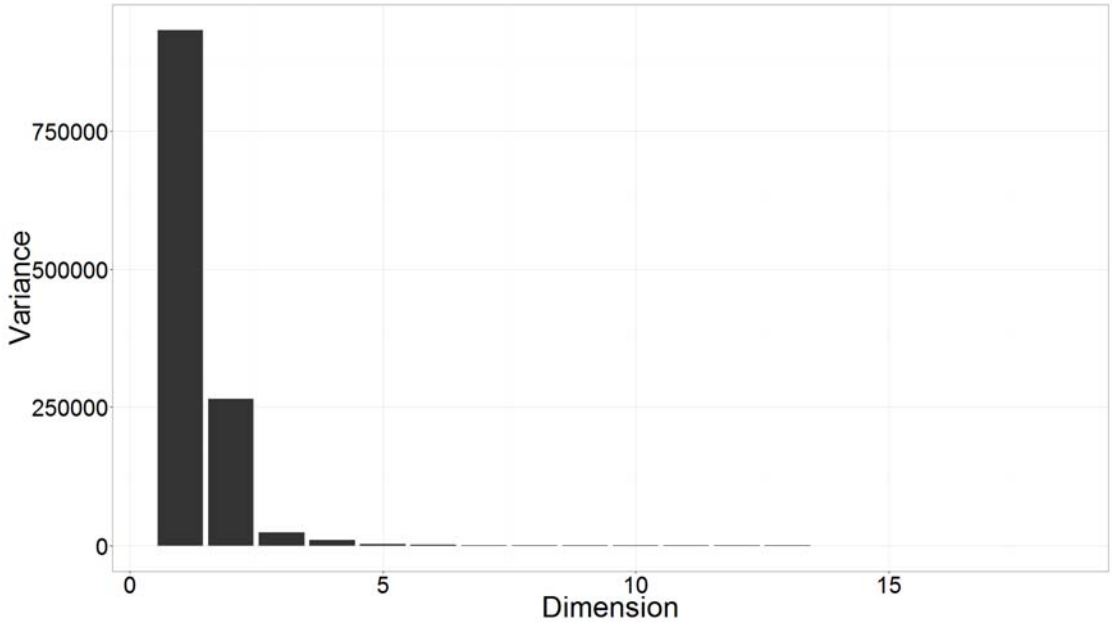


Figure 5.10 Majority of Differences in Ribavirin Treatments in Fractionation
Scree plot showing the variance in each dimension of the PCA analysis of the Ribavirin and Control treated samples. The first two components clearly capture the majority of the variance.

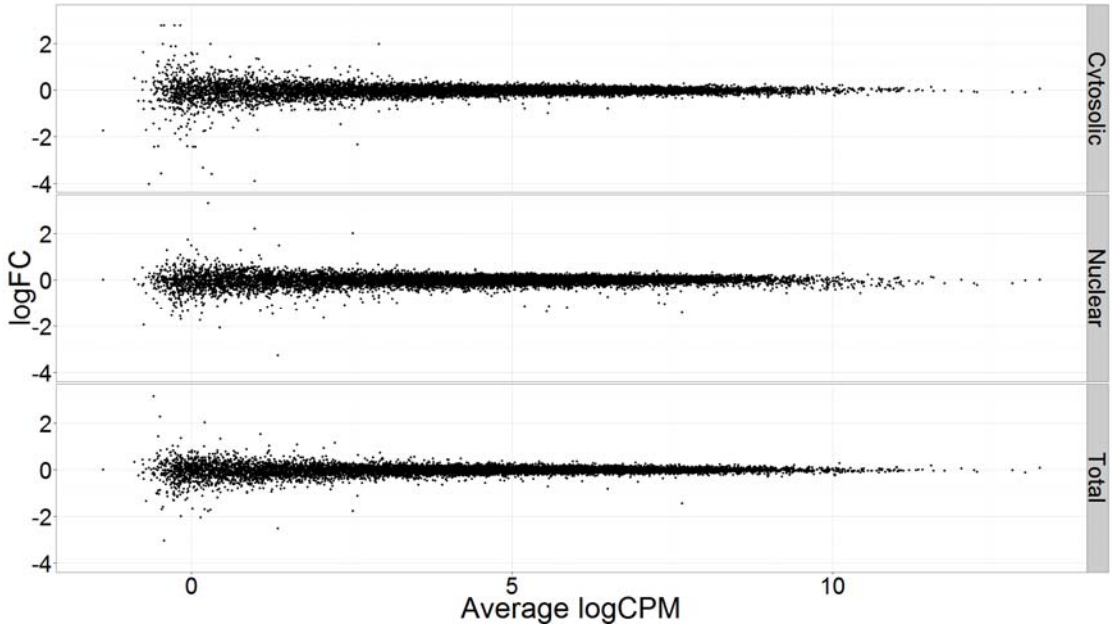


Figure 5.11 Ribavirin Smear Plot
A smear plot showing the average log counts per million (CPM) of each gene and its log fold change in the Ribavirin treatment, for each fraction. No genes were found to be differentially expressed.

5.3.3 Heat Shock vs Ribavirin RNA-Sequencing

However, the ribavirin treatment can also be viewed as the opposite of the heat shock treatment. Heat shock stimulation induces heat shock and stress response, simulating B-cell activation. Ribavirin, on the other hand, prevents the nuclear export of genes, specifically genes implicated in stress response. Comparing the RNA-sequencing data from these two treatments, in particular, shows far more up-regulation of heat and stress response genes in the heat shock samples, as expected, depicted below in Table 5.2, using GOrilla gene ontology. (Eden et al., 2007; Eden et al., 2009)

Table 5.2: Gene Ontology Pathway Enrichment in Heat Shock vs Ribavirin

Description	**P-value**	**FDR q-value**
response to unfolded protein	1.07E-16	1.23E-12
response to topologically incorrect protein	1.00E-15	5.80E-12
protein folding	3.02E-14	1.16E-10
protein refolding	7.90E-14	2.28E-10
negative regulation of inclusion body assembly	5.24E-09	1.21E-05
response to heat	8.79E-09	1.69E-05
response to temperature stimulus	2.03E-08	3.35E-05
regulation of cellular response to heat	2.12E-08	3.07E-05
cellular response to heat	3.17E-08	4.07E-05
response to chemical	3.19E-08	3.69E-05
regulation of inclusion body assembly	1.16E-07	1.22E-04
response to organic substance	3.98E-07	3.83E-04
chaperone-mediated protein folding	1.01E-06	9.01E-04
alpha-amino acid metabolic process	1.33E-06	1.10E-03

5.4 MeRIP-Seq Analysis

5.4.1 MeRIPPeR Peak Calling and Quality Control Metrics

The number of peaks called using MeRIPPeR is depicted in Figure 5.12. The control and ribavirin samples are consistent with previous studies that higher

concentrations of m^6A can be found in nuclear RNA, with the exception of the heat shock sample, which could be attributed to the loss of one of the nuclear MeRIP sample replicates. The peak enrichment was calculated by dividing the number of normalized MeRIP reads by the number of control RNA-seq reads for each sample replicate. Replicates are normalized by the total number of reads mapped, to account for differences in sequencing depth, and using the trimmed mean method (TMM) (Robinson and Oshlack, 2010) from edgeR (Robinson et al., 2010) to account for differences in the number of expressed fragments in each sample.[4]

The density distribution of log 2 normalized peak enrichment scores for peaks present in at least six out of the nine (two-thirds) of the samples is shown in Figure 5.13, which shows most of the MeRIP samples had similar IP efficiencies. However, the third cytosolic heat shock replicate and the first nuclear heat shock replicate show some major batch effects, with the third cytosolic heat shock replicate showing many peaks with low peak enrichment scores indicating a poorer IP pulldown efficiency. Figure 5.14 depicts a violin plot of the same enrichment scores. The density plot is useful for examining technical variance in the IP and the violin plot is better for visualizing dramatic shifts in the mean global IP efficiency. TMM scaling adjusts for some of the technical variance in the data.

[4] TMM scaling was calculated independently for the MeRIP and control samples to prevent scaling MeRIP samples to RNA-seq counts.

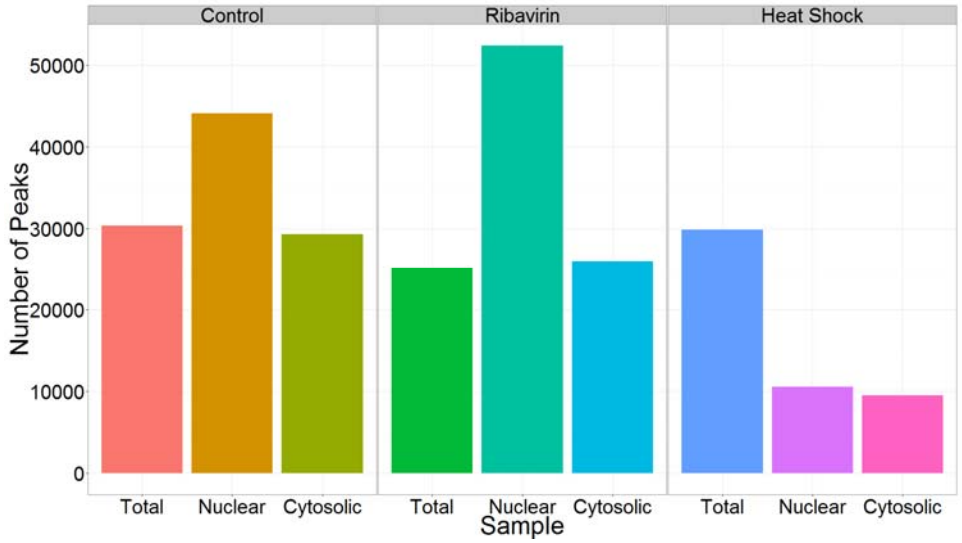


Figure 5.12 Increased Peaks Found in Nuclear Samples
The number of peaks called varies with respect to the sample, depending on the amount of m^6A present, the efficiency of the IP, and other factors. More peaks were called in the nuclear fraction, with the exception of the heat shock sample that had fewer peaks called overall.

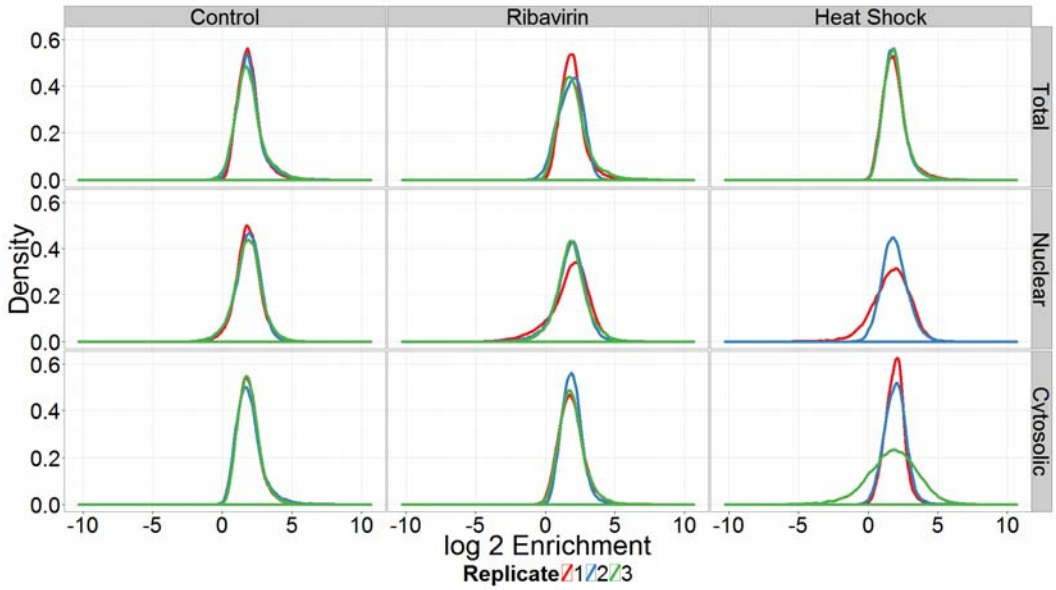


Figure 5.13 Variation in Peak Enrichment Density
The density distribution of the $\log_2$ peak enrichment of the peaks found in at least six of the nine different samples. Most of the replicates show a consistent distribution after TMM scaling, but some replicates show low IP efficiency.

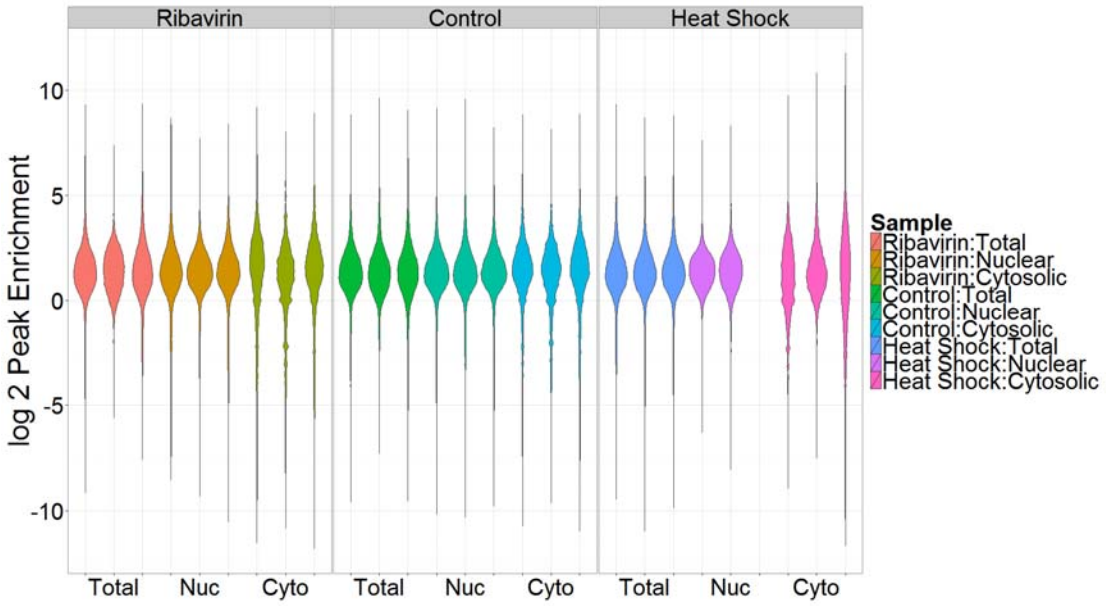


Figure 5.14 Adjusted Peak Enrichments Normalize for Technical Variation
A violin plot shows the same data as the density plot, but the density plot is useful for determining technical and biological variance, while the violin plot shows the mean IP efficiency more clearly. All of the samples have roughly the same IP efficiency, once normalized by sequencing depth and TMM scaling, though the high variance in replicates can still be observed.

The samples show similar distributions with respect to each other in the metagene plot of the peaks called, depicted in Figure 5.15. The plot is very sensitive to the number of peaks called, with fewer peaks called resulting in increased noise and a smaller signal in the 5' UTR. The figure recapitulates the enrichment of peaks found at the stop codon, as well as a smaller enrichment present in the first coding sequence exon. Principal component analysis (PCA) of the peak enrichment scores is shown in Figure 5.16 and its corresponding scree plot depicting the dimensional variances in Figure 5.17. The first dimension itself captures nearly 64% of the variance, and shows the samples clustering together with those replicates showing batch effects as outliers. The

peak enrichments do not cluster in a meaningful pattern, especially with regards to replicates, which could complicate downstream analysis.

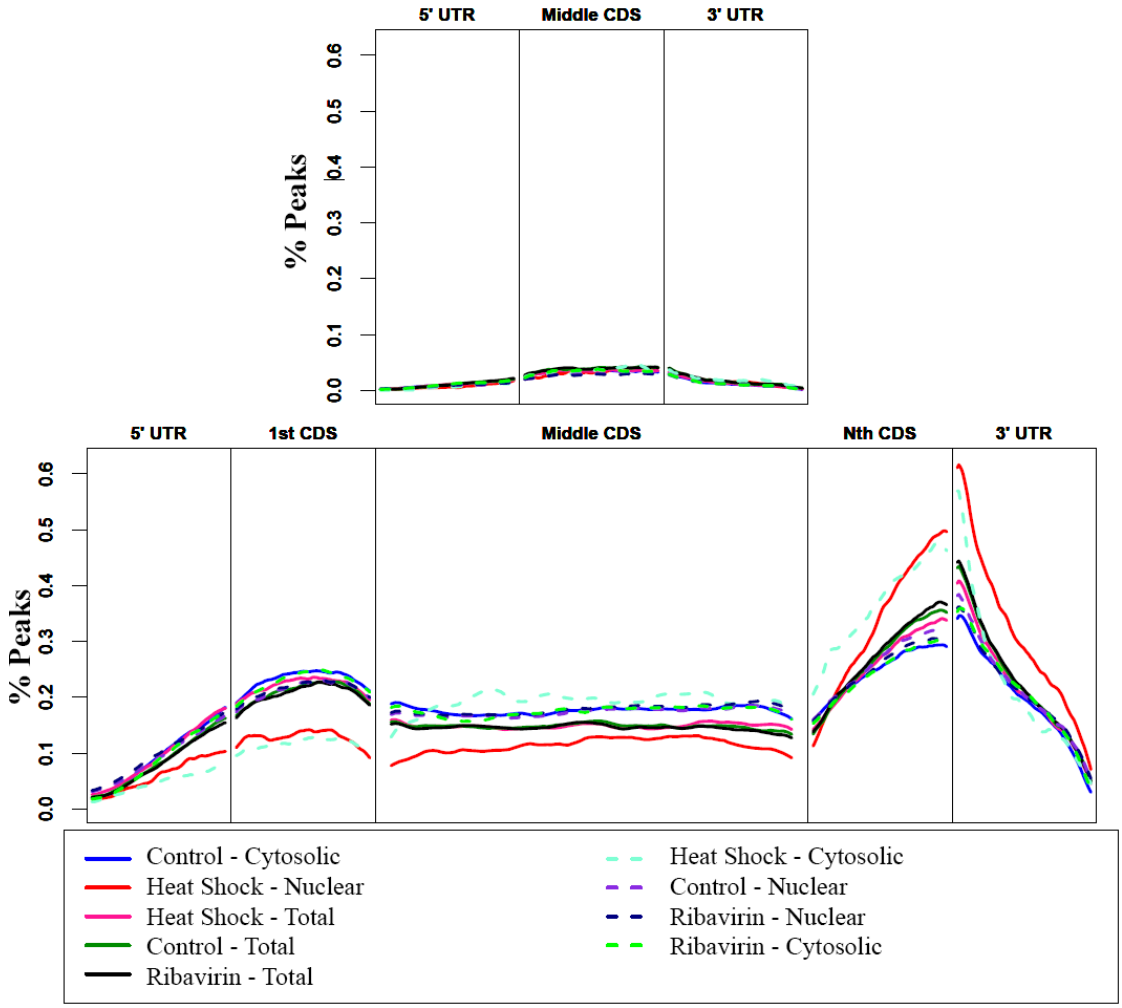


Figure 5.15 Peak Enrichment at Stop Codon and in First CDS
The metagene plot shows the distribution of peaks across a meta-genebody, with the 5' UTR and 3' UTR segments plotted separately. This particular version plots the first and last exon separately (bottom), with one and two-exon genes being plotted separately (top).

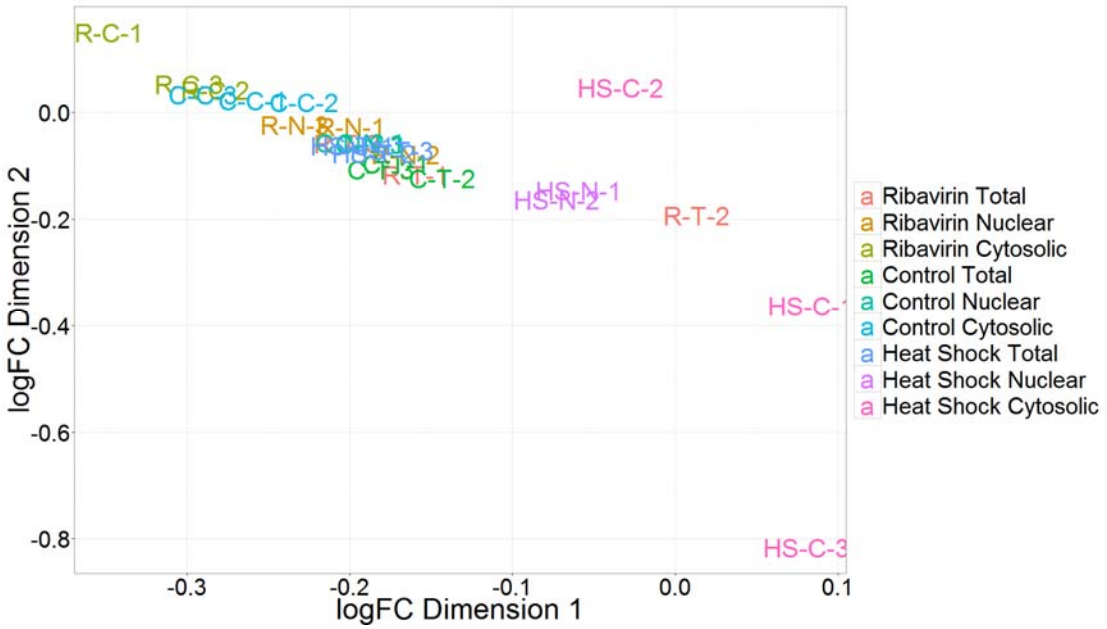


Figure 5.16 PCA of Peak Enrichments in Ribavirin/Heat Shock Samples
Principal Component Analysis (PCA) shows separation along the first dimension of samples, but the sample replicates do not cluster as well together, indicating poor IP replicability.

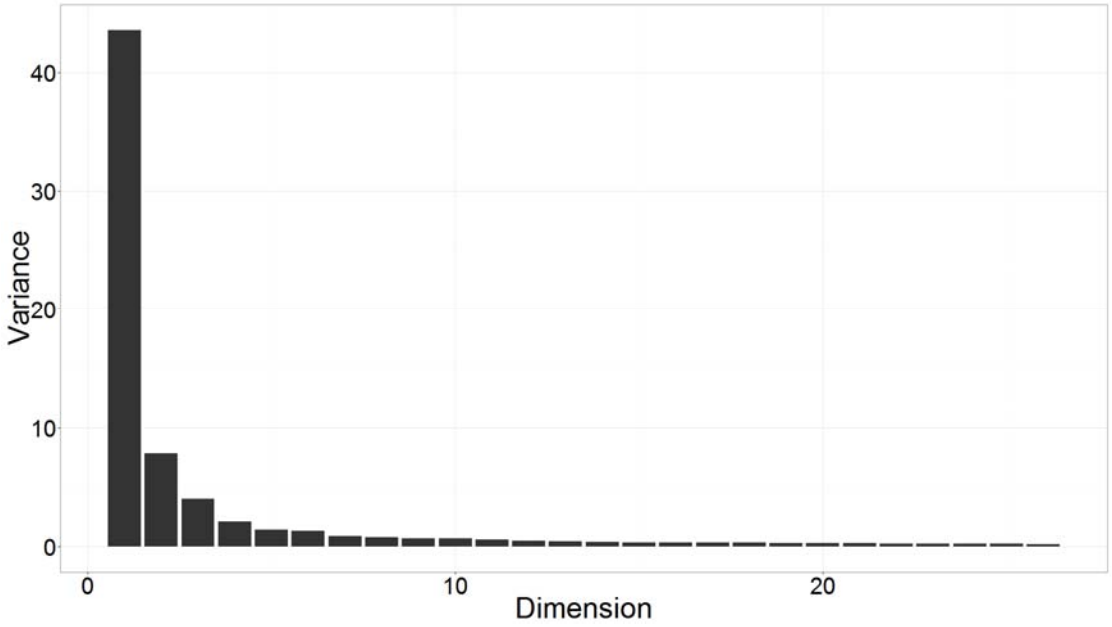


Figure 5.17 First Dimension Captures Majority of Variance in Heat Shock PCA
A scree plot of the PCA analysis of the peak enrichment in the Ribavirin and heat shock samples shows most of the variance in the first dimension.

5.4.2 Differentially Methylated Peak Regions in Heat Shock

Differentially methylated peak regions (DPMRs) were identified using the methods defined earlier in Chapter 4 Differentially Methylated Peak Regions (DMPRs), with a 100 base pair sliding windows stepped at 25 base pairs across the union of all peaks, using edgeR to analyze the count data and TMM scaling to adjust for IP efficiency and RNA-sequencing library sizes. (Quinlan and Hall, 2010; Robinson et al., 2010; Robinson and Oshlack, 2010) No peaks regions were identified to be differentially methylated in the heat shock total RNA samples, but a large subset of windows were identified as differentially methylated in the cytosolic and nuclear fractions, shown below in volcano plots in Figure 5.18 and Figure 5.19, respectively.

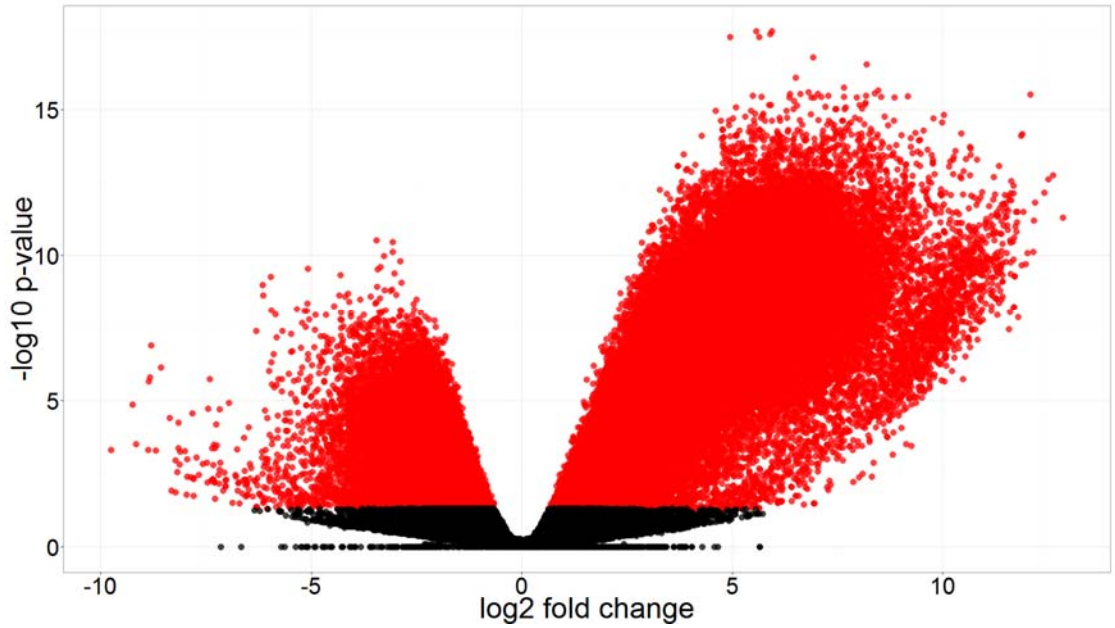


Figure 5.18 Volcano Plot of Differentially Methylated Windows in Heat Shock Cytosolic RNA
A volcano plot comparing the log 2 fold change to and the Benjamini-Hochberg adjusted -log 10 p-value of peak windows in the heat shock cytosolic RNA samples compared to the control cytosolic samples. Differentially methylated regions are denoted in red.

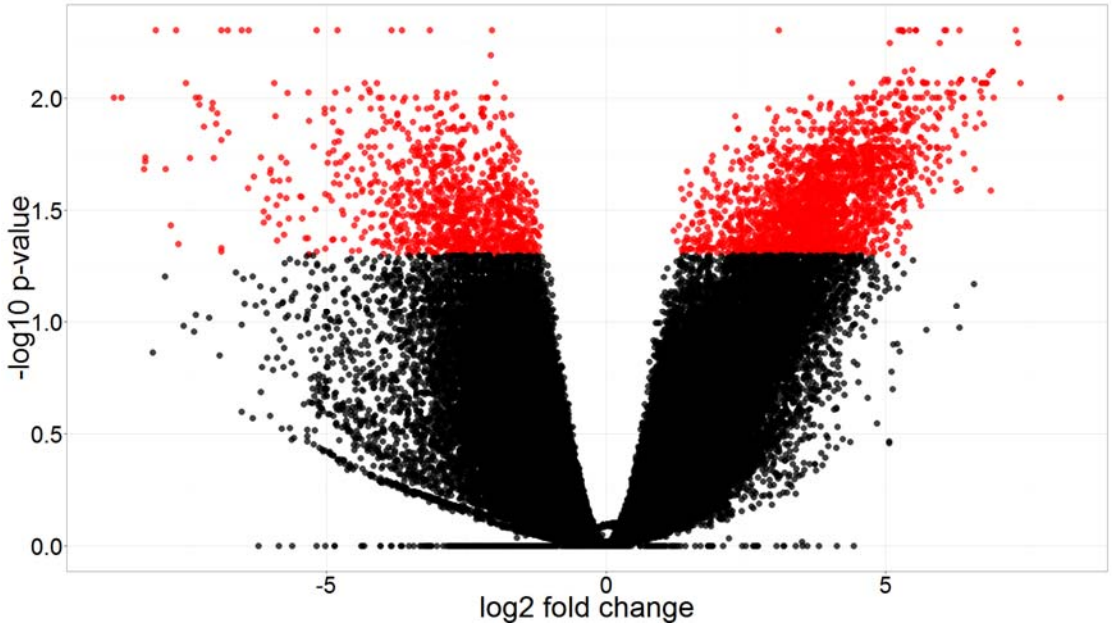


Figure 5.19 Volcano Plot of Differentially Methylated Windows in Heat Shock Nuclear RNA
A volcano plot comparing the log 2 fold change to and the Benjamini-Hochberg adjusted -log 10 p-value of peak windows in the heat shock nuclear RNA samples compared to the control cytosolic samples. Differentially methylated regions are denoted in red. Fewer windows are differentially methylated relative to the cytosolic fraction.

Annotating these regions to genes, the metagene distribution of the DMPRs, plotted separately for hypermethylated and hypomethylated regions, is shown in Figure 5.20 for the cytosolic fraction and Figure 5.21 for the nuclear fraction. Identifying fewer DMPRs results in a noisier signal, such as the hypermethylated regions in the heat shock cytosolic vs nuclear fractions. However, there does appear to be a strong signal of hypomethylation in the cytosol and hypermethylation in the nucleus around the stop codon, as well as hypomethylation in the 5' UTR and first coding sequence exon in the cytosol and nucleus.

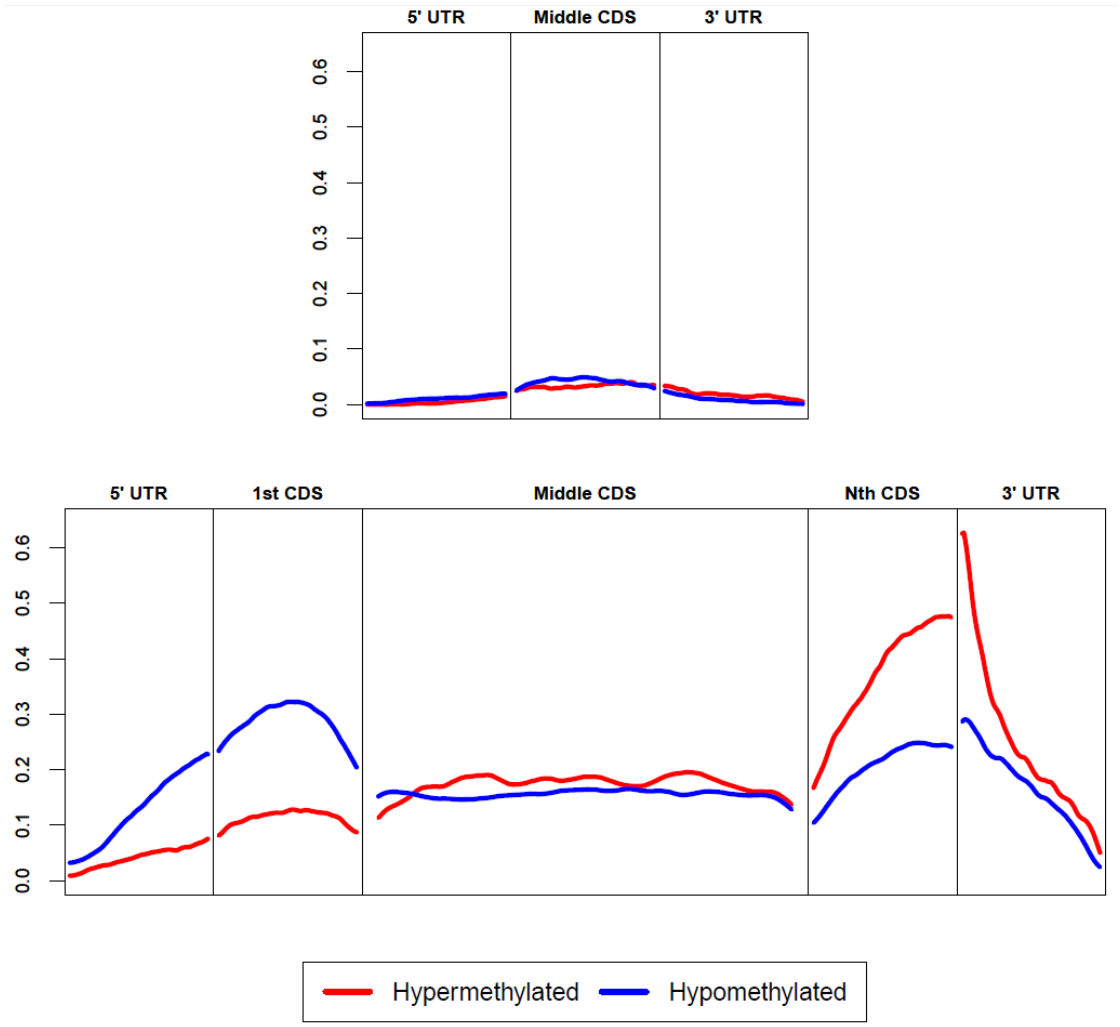


Figure 5.20 Heat Shock DMPRs in Cytosol shows Hypomethylation in First CDS, Hypermethylation at Stop Codon

A binned metagene shows the distribution of differentially methylated peak regions mapped to gene features, with the first and last coding exons plotted separately. Genes with two and fewer exons are plotted separately on top in the traditional metagene plot. Heat Shock DMPRs in the Cytosolic fraction show a strong hypomethylation signal in the first coding sequence and 5' UTR and a hypermethylation signal at the stop codon.

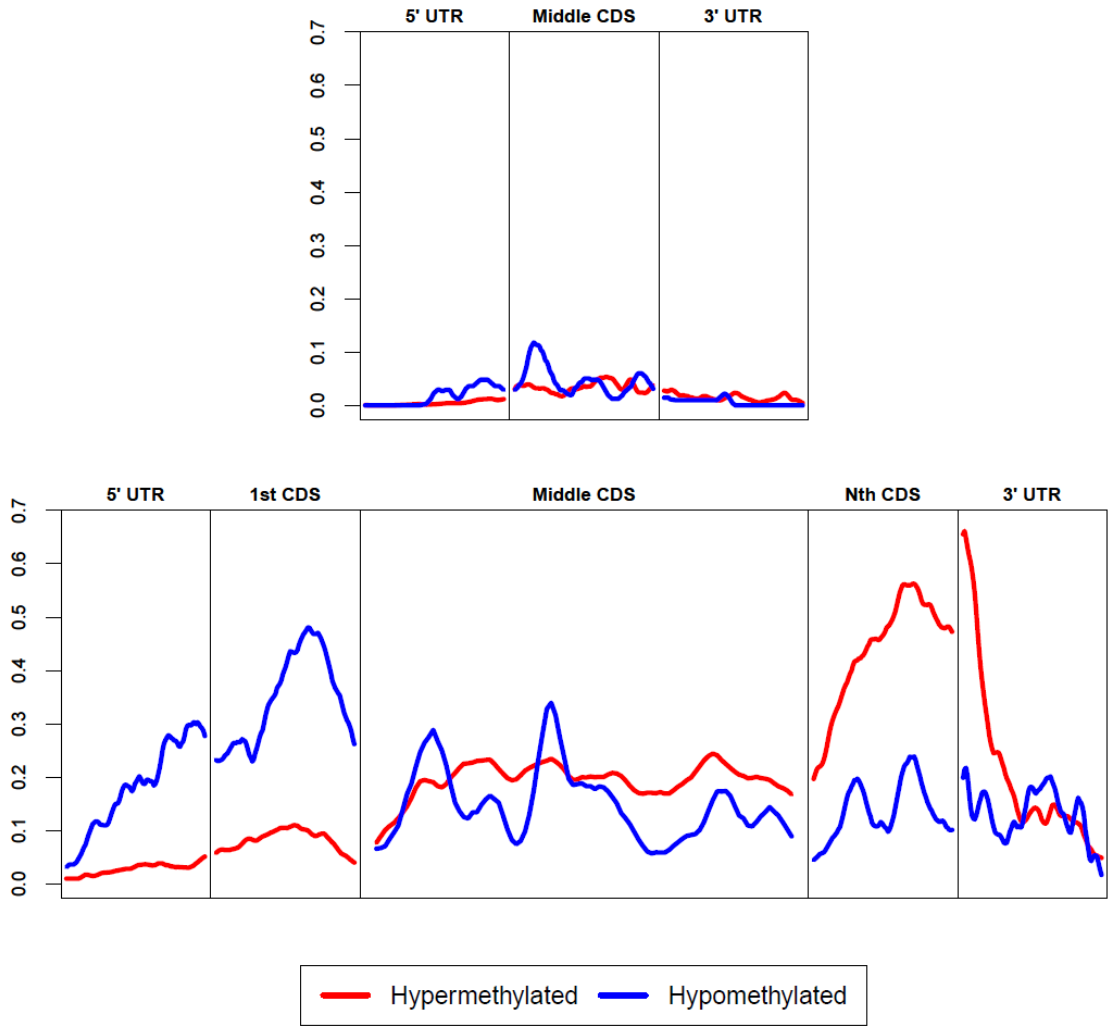


Figure 5.21 Heat Shock DMPRs in Nucleus shows Hypomethylation in First CDS, Hypermethylation at Stop Codon

A binned metagene shows the distribution of differentially methylated peak regions mapped to gene features, with the first and last coding exons plotted separately. Genes with two and fewer exons are plotted separately on top in the traditional metagene plot. Heat Shock DMPRs in the Nuclear fraction show a strong hypomethylation signal in the first coding sequence and 5' UTR and a hypermethylation signal at the stop codon. With fewer DMPRs in the nucleus fraction than in the cytosolic, the level of noise is increased.

To determine if there is in fact a correlation between a type of hyper- or hypo-methylation and the change in nuclear export, Figure 5.22 shows the distribution of the log fold change in the nuclear to cytosolic ratio in the heat shock samples for each gene feature and change in direction, separately. Most of the distributions are centered on 0, indicating no correlation, with the exception of differentially methylated peaks in intronic regions. Hypomethylated DMPRs in introns are correlated with a decrease in the mRNA nuclear to cytosolic ratio, while hypermethylated DMPRs in introns show the opposite.

In other words, the change in methylation in introns is correlated with the change in the mRNA nuclear/cytosolic fraction. However, DMPRs in the nuclear/cytosolic ratio are dependent on the mRNA nuclear/cytosolic fraction, to account for changes in mRNA levels to determine the change in methylation. A very small shift, but not statistically significant, can be observed in the other gene features for the nuclear/cytosolic DMPRs, but none of them are as significant as the change in the intronic DMPRs, indicating that this correlation is not because of the dependency.

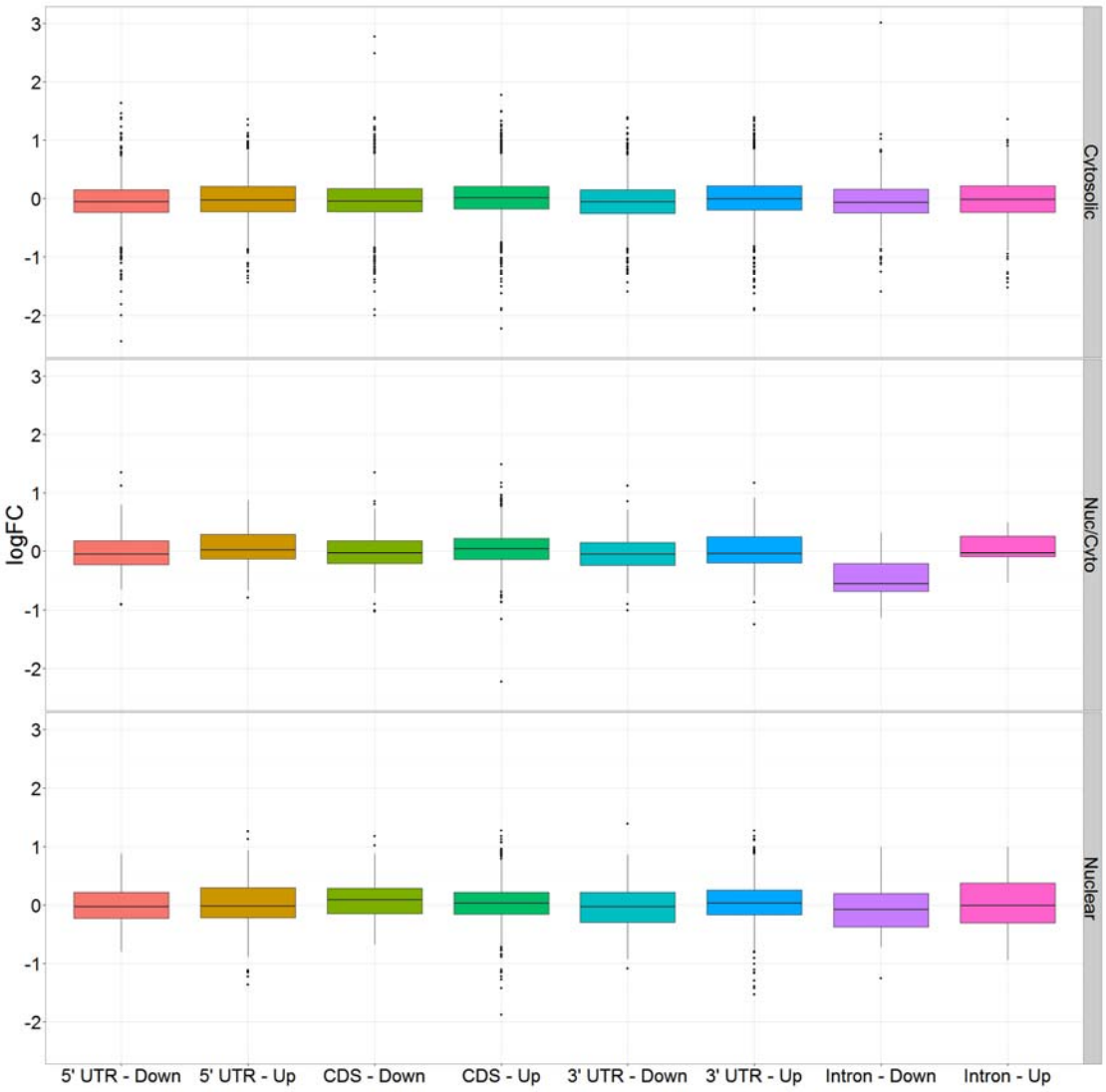


Figure 5.22 Boxplot of Heat Shock Log Fold Change in Nuclear/Cytosolic RNA-Seq Ratio by Heat Shock DMPR Gene Annotations
The x-axis is the gene feature to which a DMPR was annotated and the y-axis is the log fold change of the nuclear/cytosolic ratio in the RNA-sequencing data analyzed earlier, separated by the DMPR type and direction. Most of the distributions are relatively the same, with the means still remaining between the quartiles, with the exception of hypomethylation in the intron correlated with decreased nuclear/cytosolic log fold change.

5.4.3 Differentially Methylated Peak Regions in Ribavirin Treatment

The Ribavirin treated samples do not show as significant changes as the heat shock samples, both in the RNA-sequencing, as well as in the IP. No DMPRs

were found in either the Ribavirin total, cytosolic, or nuclear fractions. The Ribavirin nuclear/cytosolic DMPR analysis did, however, produce many DMPRs, as shown in Figure 5.23. The metagene distribution of these DMPRs is shown in Figure 5.24, showing an enrichment of hypomethylation at the stop codon. The boxplot comparison of annotated gene features to the nuclear/cytosolic log fold change, similar to Figure 5.22 discussed earlier, in Figure 5.25 does not recapitulate the results from 5.4.2 Differentially Methylated Peak Regions in Heat Shock.

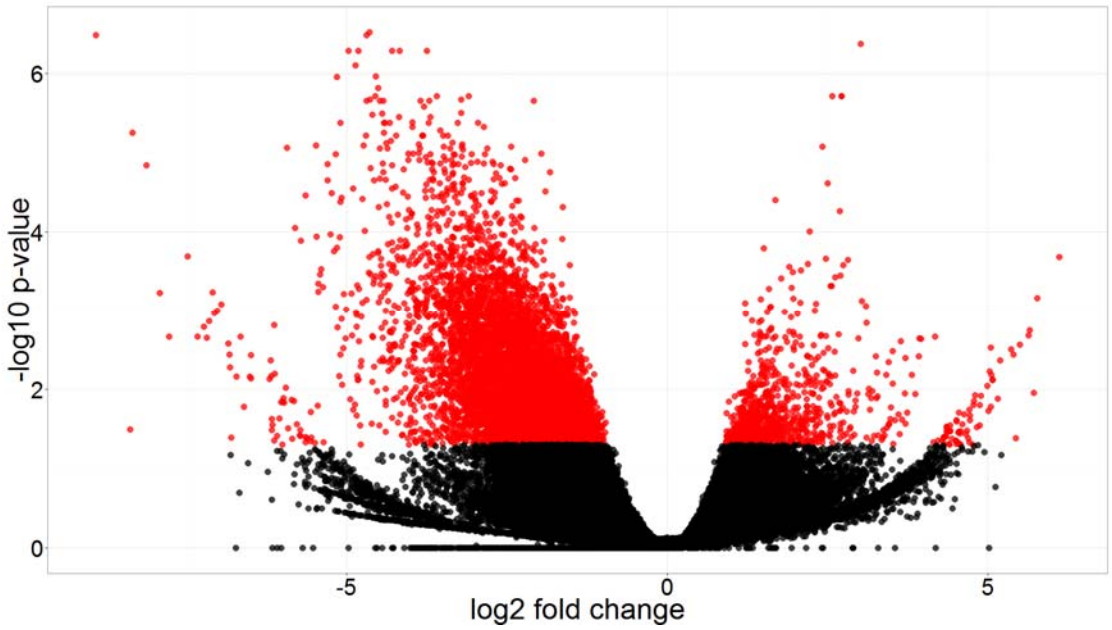


Figure 5.23 Volcano Plot of Differentially Methylated Windows in Ribavirin Cytosolic vs Nuclear Fractions
A volcano plot comparing the log 2 fold change to and the Benjamini-Hochberg adjusted -log 10 p-value of peak windows in the Ribavirin treated samples nuclear to cytosolic fractions compared to that of the control cytosolic samples. Differentially methylated regions are denoted in red.

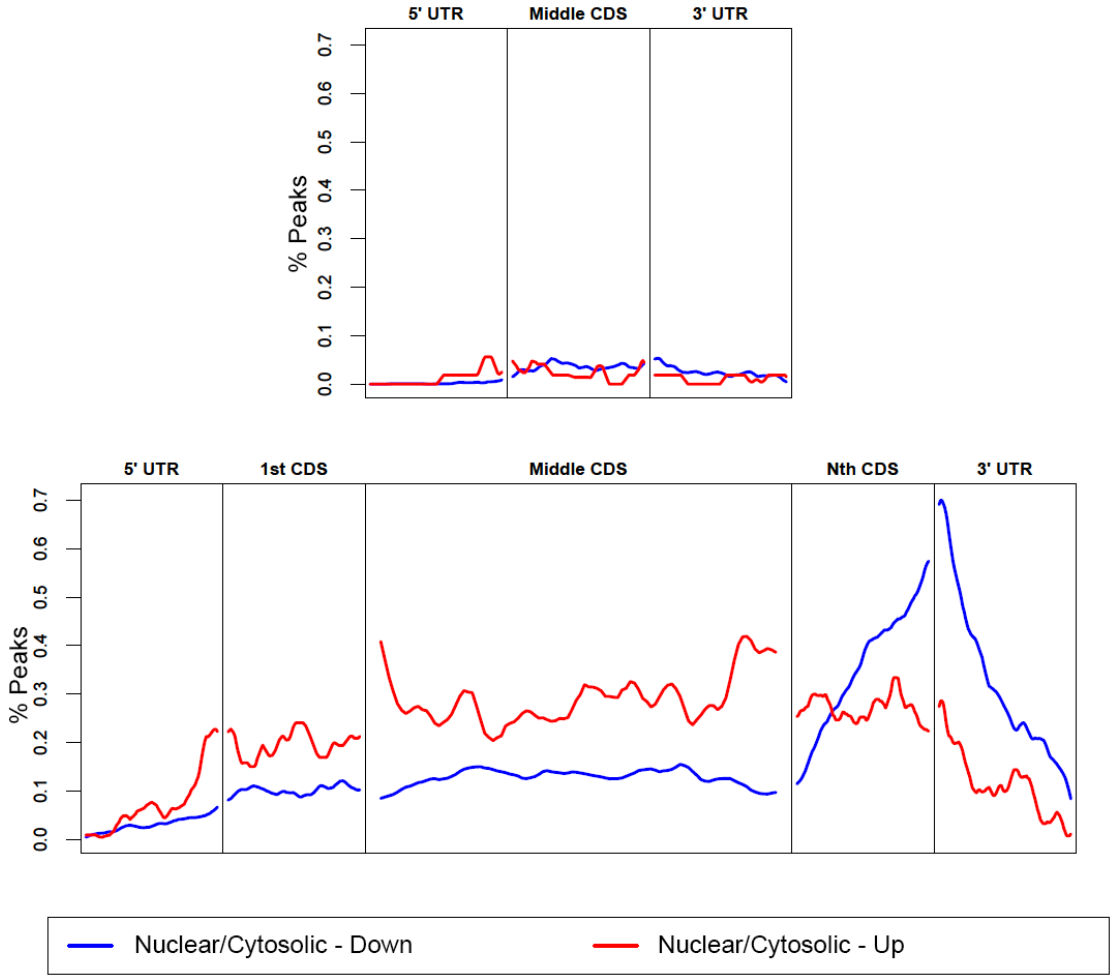


Figure 5.24 Metagene plot of Differentially Methylated Windows in Ribavirin
Metagene plot showing the profile of differentially methylated peak window regions in the Ribavirin treatment cytosolic versus nuclear fraction. Hypomethylation is enriched at the stop codon and in the 3' UTR, while hypermethylation doesn't have a consistent metagene profile.

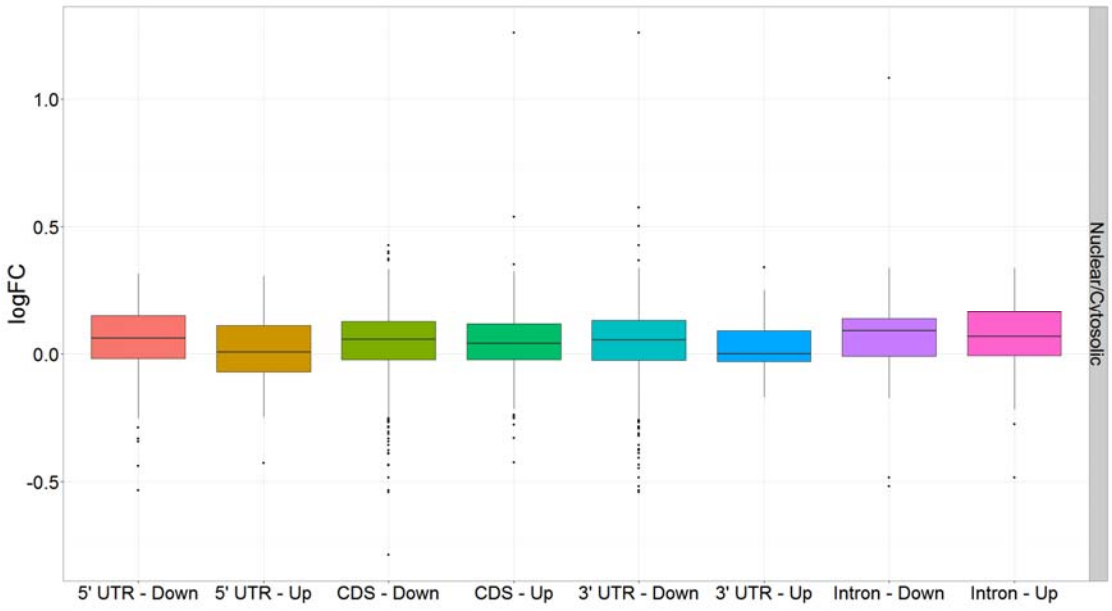


Figure 5.25 Boxplot of Ribavirin Log Fold Change in Nuclear/Cytosolic RNA-Seq Ratio by Ribavirin DMPR Gene Annotations
The x-axis is the gene feature to which a DMPR was annotated and the y-axis is the log fold change of the nuclear/cytosolic ratio in the RNA-sequencing data analyzed earlier, separated by the DMPR type and direction. This distribution does not recapitulate the pattern observed earlier in Figure 5.22.

5.4.4 Differentially Methylated Peak Regions in Heat Shock vs Ribavirin

Previously, RNA-sequencing data was compared between the heat shock and ribavirin treatments, and similar comparisons can be made with the differentially methylated peak regions. Depicted in Figure 5.26, the volcano plot shows most of the regions are hypomethylated in the heat shock sample, with most of these regions mapping to the stop codon and 3' UTR. Table 5.3 illustrates the pathway enrichment for the genes that these regions map to, which includes many RNA regulatory and splicing pathways. This demonstrates that m^6A has the potential to be directly involved in these regulatory pathways, further corroborated by similar results in the nuclear fraction.

Table 5.3: Gene Ontology Pathway Enrichment Genes with Differentially Methylated Peak Regions in Heat Shock vs Ribavirin in Total RNA

Description	P-value	FDR q-value
nucleic acid metabolic process	3.51E-26	4.05E-22
RNA metabolic process	2.34E-24	1.35E-20
cellular macromolecule metabolic process	1.26E-23	4.86E-20
nucleobase-containing compound metabolic process	2.8E-23	8.1E-20
heterocycle metabolic process	5.97E-22	1.38E-18
cellular aromatic compound metabolic process	9.67E-22	1.86E-18
organic cyclic compound metabolic process	7.2E-21	1.19E-17
regulation of gene expression	2.04E-20	2.95E-17
regulation of cellular macromolecule biosynthetic process	3.34E-20	4.29E-17
regulation of macromolecule biosynthetic process	7.34E-20	8.48E-17
macromolecule metabolic process	1.3E-19	1.36E-16
regulation of biosynthetic process	5.75E-19	5.53E-16
regulation of cellular biosynthetic process	9.35E-19	8.31E-16
macromolecule biosynthetic process	7.67E-18	6.33E-15
regulation of nucleobase-containing compound metabolic process	1.09E-17	8.43E-15
cellular macromolecule biosynthetic process	2.25E-17	1.63E-14
cellular nitrogen compound metabolic process	4.17E-17	2.84E-14
regulation of metabolic process	5.53E-17	3.55E-14
regulation of cellular metabolic process	6.27E-17	3.81E-14
regulation of macromolecule metabolic process	6.93E-17	4E-14
regulation of nitrogen compound metabolic process	2.88E-16	1.58E-13
regulation of RNA metabolic process	4.72E-16	2.48E-13
regulation of transcription, DNA-templated	1.12E-15	5.61E-13
regulation of primary metabolic process	2.06E-15	9.93E-13
regulation of RNA biosynthetic process	2.23E-15	1.03E-12
regulation of nucleic acid-templated transcription	2.56E-15	1.14E-12
nitrogen compound metabolic process	6.7E-15	2.87E-12
RNA biosynthetic process	8.03E-14	3.32E-11
transcription, DNA-templated	1.13E-13	4.49E-11
nucleic acid-templated transcription	1.21E-13	4.66E-11
RNA processing	1.03E-12	3.84E-10
nucleobase-containing compound biosynthetic process	3.03E-12	1.09E-09
primary metabolic process	7.2E-12	2.52E-09
RNA splicing	8.34E-12	2.83E-09
heterocycle biosynthetic process	1.68E-11	5.54E-09
aromatic compound biosynthetic process	1.74E-11	5.59E-09

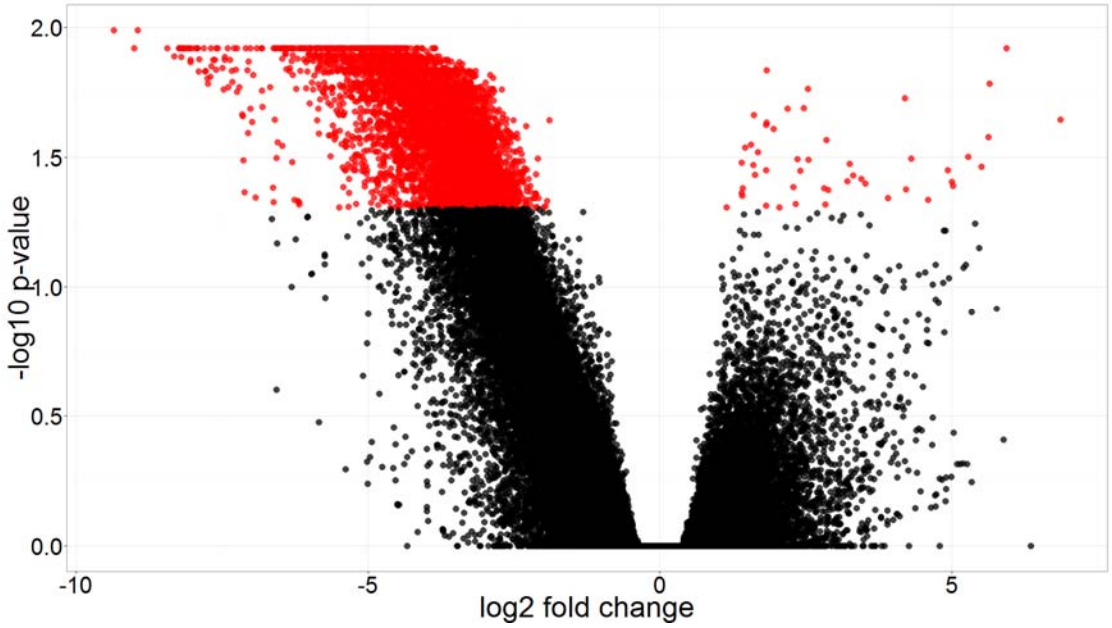


Figure 5.26 Volcano Plot of Differentially Methylated Windows in Heat Shock vs Ribavirin Total RNA
A volcano plot comparing the log 2 fold change to and the Benjamini-Hochberg adjusted -log 10 p-value of peak windows in the heat shock vs ribavirin treatment total RNA. Differentially methylated regions are denoted in red.

5.5 Discussion and Conclusions

Multiple comparisons can be made within the extensive experimental design to examine the role of heat shock and Ribavirin treatments in m^6A in the context of nuclear and cytosolic RNAs. The heat shock treatments induced a dramatic response, as expected, with many genes up-regulated in pathways responding to temperature and heat response, including genes in the HSP70 gene family. The ribavirin treatments serve as the opposite to the heat shock, restricting the nuclear export of genes particularly involved in stress response. Although, the ribavirin treatments did not significantly affect the RNA-sequencing expression levels of genes relative to the control samples.

Differentially methylated peaks in the heat shock samples were correlated with changes in the nuclear to cytosolic ratio, indicating that m^6A could play a role in

nuclear export of mRNAs. Comparing these results with EIF4E IPs, for example, could help in connecting the pathways. Ribavirin is known to affect the nuclear export of specific genes, such as BCL6 and BCL2,[5] but these genes were not significantly impacted in the RNA-sequencing data. The original results were based on qPCR data, and RNA-sequencing often have higher variance, confounding analysis.

Heat shock stimulation is a widely-studied form of stress response, and specifically used in B-cell cell lines to simulate B-cell activation. Ribavirin interacts with EIF4E, preventing the nuclear export of many of the stress response genes up-regulated in heat shock treatment. Comparing the two treatments, the RNA-sequencing data showed a strong enrichment for genes involved in heat and stress response pathways. The MeRIP-Seq data showed a strong hypomethylation signal, primarily near the stop codon and 3' UTR, in genes corresponding to RNA regulatory pathways, including splicing. While YTHDF2 has been studied as an m^6A reader in the cytosol (Wang et al., 2014a), this does not explain the function of m^6A in the nucleus. The treatment data and comparisons show that m^6A sites change dramatically in genes responsible for RNA regulation, export, and splicing. Identifying nuclear-specific readers could help further elucidate its exact functional role.

[5] Unpublished data from Leandro Cerchietti, MD and Katherine Borden, PhD.

CHAPTER 6 THE ROLE OF METHYL-6-ADENOSINE IN ADIPOGENESIS: A CASE STUDY IN PORCINE MODEL

6.1 Introduction

The fat mass and obesity associated FTO gene was discovered to be one of the demethylases of methyl-6-adenosine. (Jia et al., 2011) Similar in structure to the AlkB family of DNA demethylases, (Gerken et al., 2007) FTO is also an alpha-ketoglutarate and iron (II) dependent dioxygenase. Specific alleles of the FTO gene were found to be correlated with obesity in humans, (Frayling et al., 2007; Yang et al., 2012) further implicating m^6A in adipogenesis. These implications were confirmed by further experiments in mouse cell lines, where FTO levels were found to decrease and m^6A content increased as fat cells matured. (Zhao et al., 2014)

Although m^6A has not yet been profiled in the porcine model, GWAS studies have found similar correlations with FTO and fat mass across different pig breeds. (Fan et al., 2009; Fontanesi et al., 2010; Fontanesi et al., 2009) The Jinhua breed of pigs is native to China, with superior meat quality and slow muscle growth compared to the faster-growing and leaner Danish breed, Landrace. (Miao et al., 2009) Jinhua pigs exhibit an intramuscular fat (IMF) content of around 4.54%, compared to 1.43% in Landrace pigs, (Guo et al., 2011) which not only explains the higher meat quality, but can be used as the perfect model to study m^6A in the context of adipogenesis. In addition, global m^6A levels have been found to be anti-correlated with adipose levels, with increased expression of FTO and decreased m^6A levels in samples with higher tri-glyceride and fat content than those with lower.[6]

[6] Unpublished data in collaboration with Xinxia Wang, PhD and Yizhen Wang, PhD.

6.2 Methods

Tissue samples from three different biological replicates of Jinhua and Landrace pigs, each, were taken from the soleus and tibialis anterior muscle, in collaboration with Qing Wu Shen, PhD, Xinxia Wang, PhD, and Yizhen Wang, PhD. RNA was extracted and polyA-purified, and a single round of MeRIP-Seq was performed on the samples. Following Illumina TruSeq Stranded RNA library preparation, libraries were sequenced on four lanes at single-ended 50 base pairs on an Illumina HiSeq 2500 sequencer.

6.3 Results

6.3.1 Distribution of Reads and m^6A Peaks

The distribution of reads sequenced is shown in Figure 6.1, which shows a variable amount of ribosomal contamination and decreased 5' UTR representation, which could be indicative of both poor polyA-purification and RNA quality. The MeRIP samples also do not show as dramatic of a shift in read distributions compared to other samples, which could be due to poorer IP efficiency but most likely the lower sequencing depth. The number of peaks called by MeRIPPeR is shown in Figure 6.2, with relatively few peaks being called compared to previous studies, most likely due to the low sequencing depth. The distribution of these peaks to gene features is shown in Figure 6.3, with most peaks mapping to intergenic regions, as can be expected from the read distribution in Figure 6.1.

Peaks were called by requiring presence in at least two of three replicates, because of lack of replicates and poorer IP efficiency. The distribution of peak enrichment is depicted in Figure 6.4, which shows a lack of concordance between replicates in the Jinhua Tibialis Anterior and Landrace Soleus samples,

which partially explain the few number of peaks called in those samples (Figure 6.2). The peak enrichments were computed by taking the union of all peaks; differentially methylated peaks are likely to be found towards the maximum and minimum edges. Despite taking the union, most peak windows do still show a positive peak enrichment, indicating that few peaks show significant demethylation when comparing the samples. The metagene distribution of reads depicted in Figure 6.5 do not show significant enrichments around the 3' UTR or transcription start site (TSS).

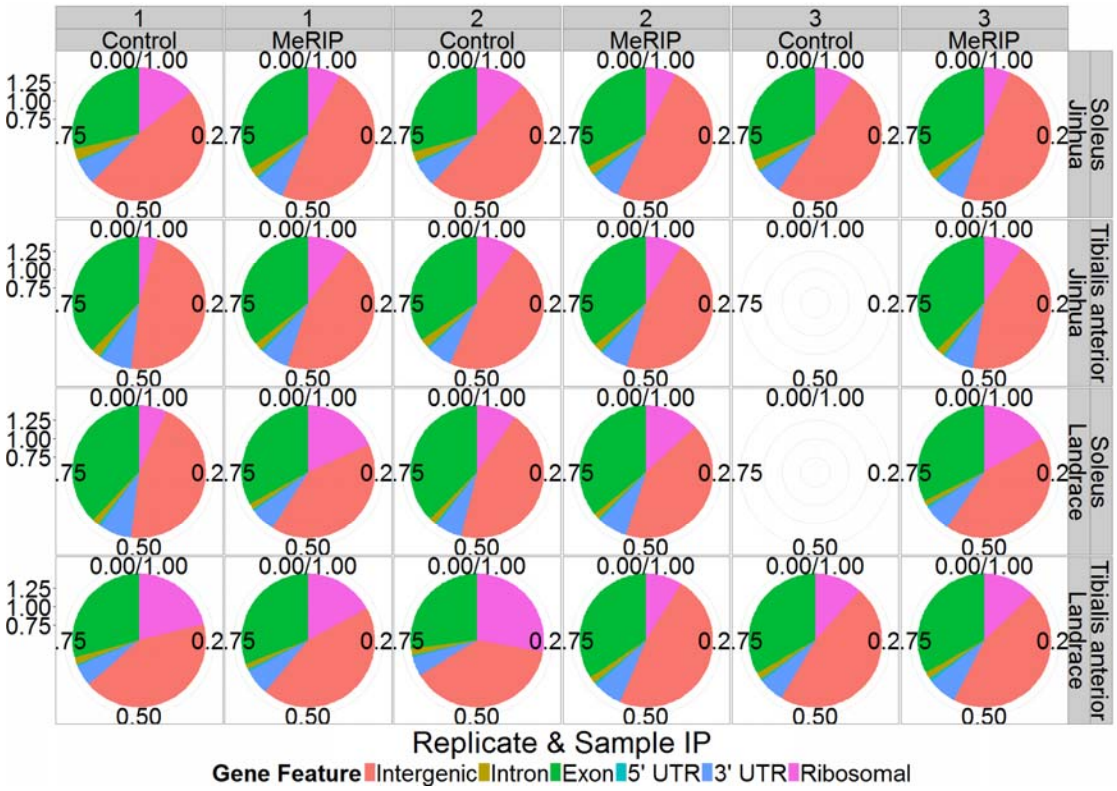


Figure 6.1 Increased Reads Mapping to Intergenic Regions
Pie charts showing the distribution of reads mapping to gene features, with intergenic in salmon, introns in dark yellow, exonic regions in dark green, 5' UTR in cyan, 3' UTR in blue, and ribosomal in pink. Most of the reads map to intergenic regions in the porcine model using RefSeq annotations, perhaps coming from genes that have yet to be annotated.

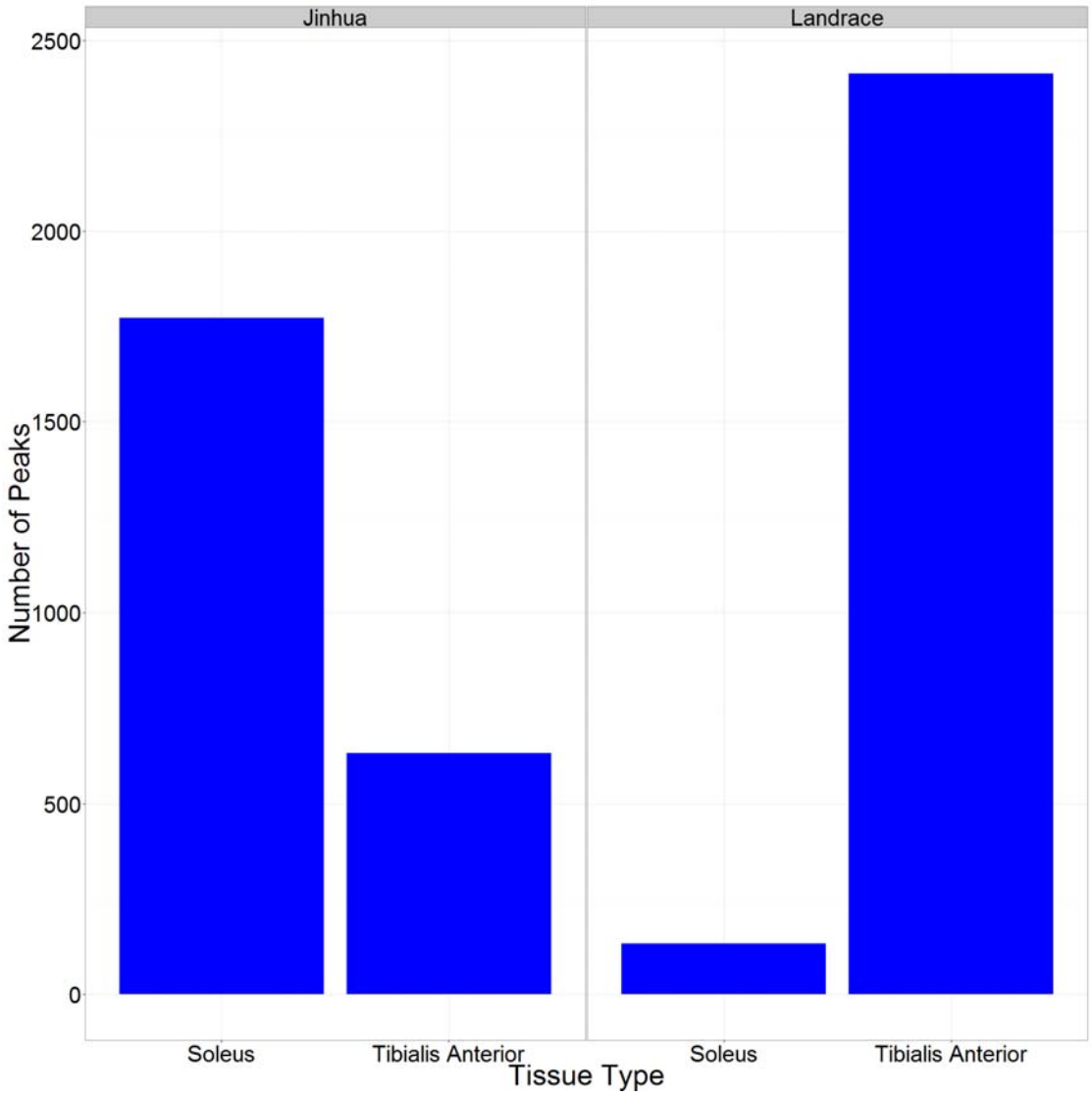


Figure 6.2 High Variation in Number of Called Peaks
Number of total peaks called per sample, after requiring replicates be present in at least two of three replicates. Relatively fewer peaks were called in the Landrace Soleus sample, with few peaks called overall compare to IPs in other species.

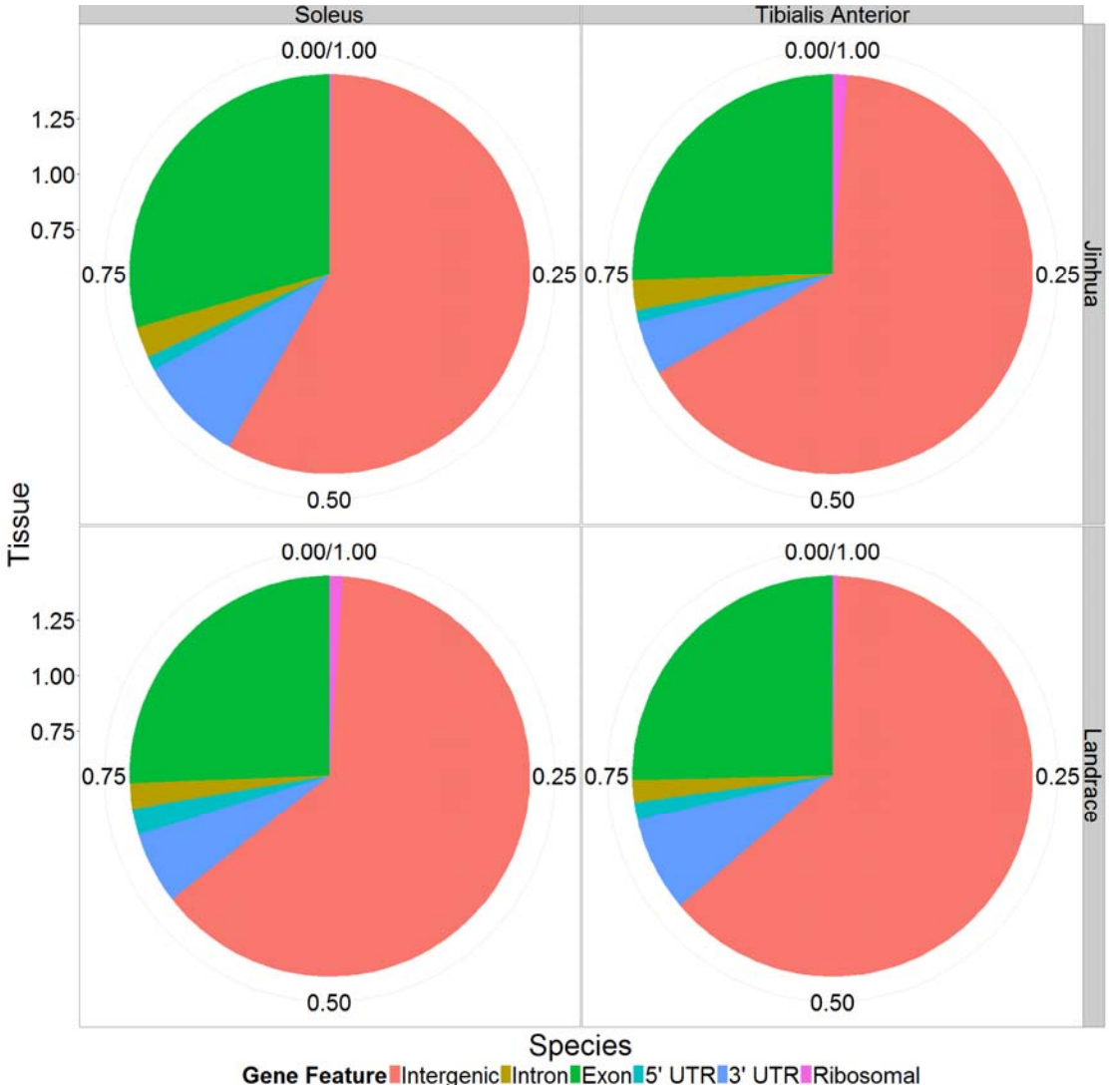


Figure 6.3 Increased Number of Peaks mapping to Intergenic Regions
Pie charts showing the distribution of peaks called mapping to gene features, with intergenic in salmon, introns in dark yellow, exonic regions in dark green, 5’ UTR in cyan, 3’ UTR in blue, and ribosomal in pink. The distribution of peaks called to gene features shows most peaks mapping to intergenic regions and very few peaks to ribosomal regions. Fewer peaks mapped to the 3’ UTR than previous studies.

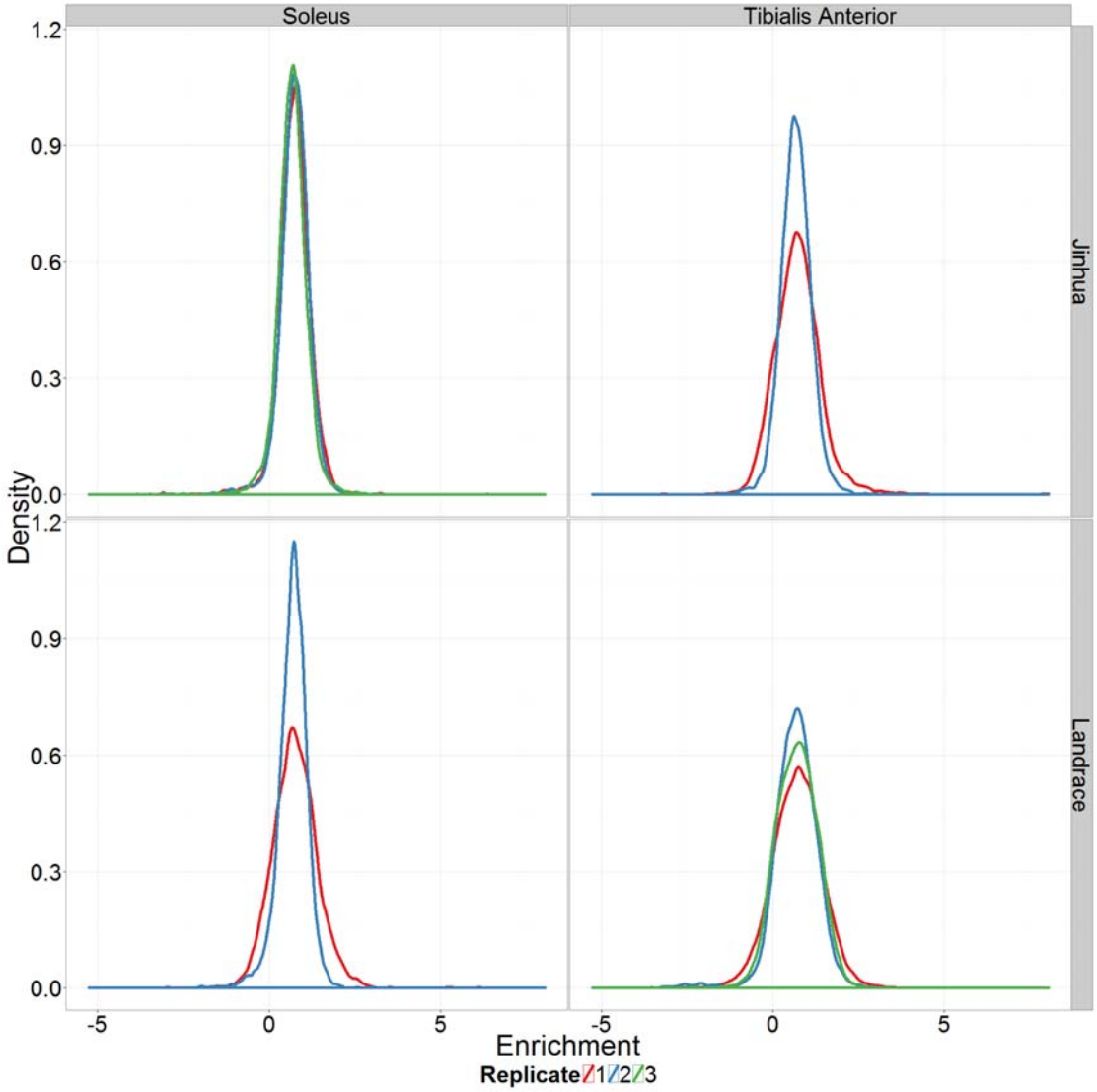


Figure 6.4 Low Replicability in Landrace Soleus Explains Low Peak Numbers
The density of peak enrichment scores for each replicate is shown, with the numbered replicates shown as red, blue, and green, respectively. Peak enrichment density shows most peaks are enriched to the same degree, indicating a successful IP. The peak distributions are most correlated in the Jinhua Soleus sample, with worse replicability the Jinhua Tibialis Anterior and Landrace Soleus samples, correlated with lower number of peaks called.

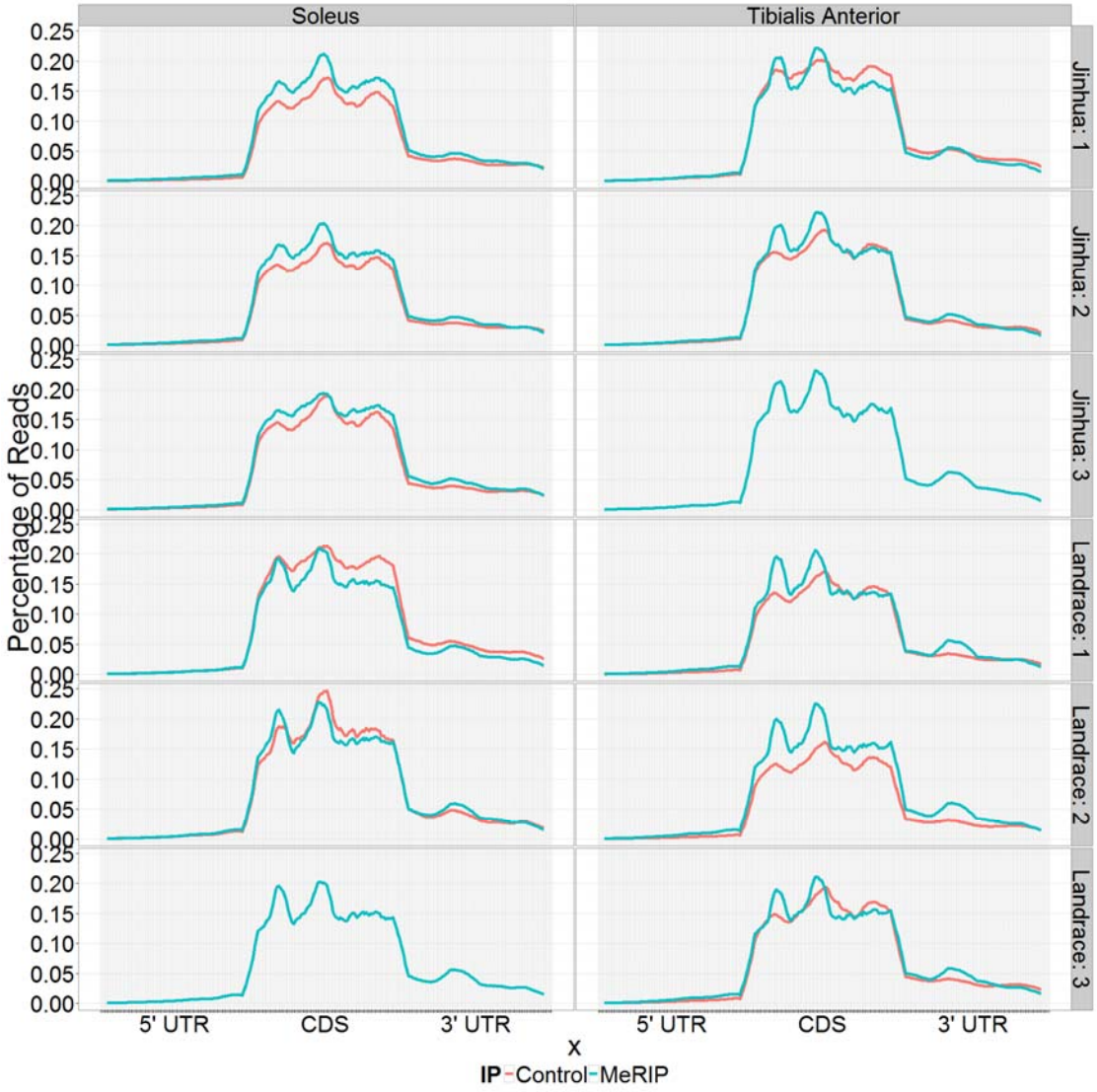


Figure 6.5 Read Metagene Affected by Lack of Annotation
Metagene distribution of reads mapped to gene features does not show a significant increased around the transcription start site or stop codon. With most peaks mapping to unannotated intergenic regions, the amount of signal present is far lower.

6.3.2 Differentially Expressed Genes

The MeRIP-seq protocol (Meyer et al., 2012) requires a normal control RNA-sequencing sample for each MeRIP-Seq sample to normalize for transcript abundances in each replicate. In a larger experimental design, such as the two

by two study in the porcine model, this data can provide additional information on differentially expressed genes between samples. A multidimensional scaling (MDS) plot that minimizes distances between the normalized gene counts, Figure 6.6, shows a clear separation in the RNA-seq data on the first dimension with respect to the species. The separation on the second-dimension shows some separation with respect to the tissue, with the exception of the Jinhua tibialis anterior replicate one.

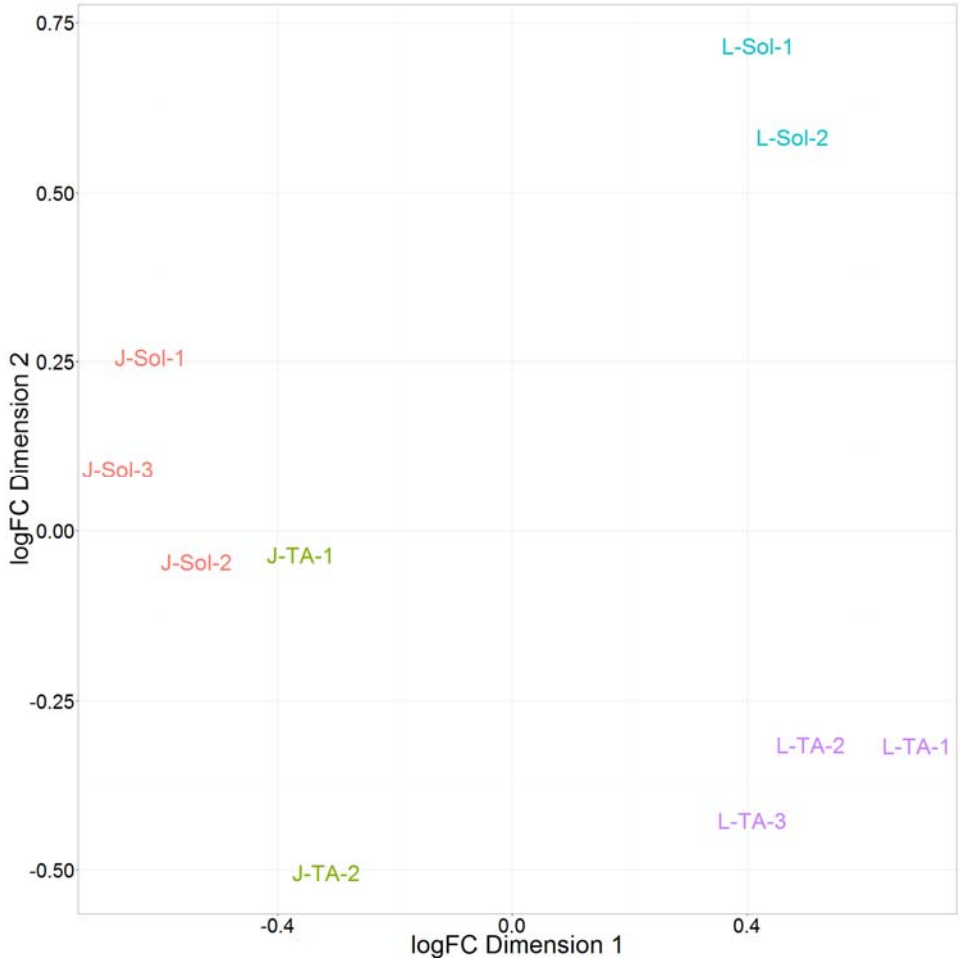


Figure 6.6 MDS Plot Clusters Samples by Species and Tissue
Multi-Dimensional Scaling (MDS) plot of the RNA-seq data shows a clear separation by species, with Jinhua and Landrace separated by the first dimension on the x-axis. The data is somewhat separated by tissue, with the exception of the Jinhua tibialis anterior replicate 1.

A volcano plot comparing the log 2 fold change and the adjusted p-value is shown below in Figure 6.7, which shows more than half of the genes appear to be differentially expressed between the two species. Volcano plots comparing the individual tissues separately produced similar results, though the focus is on comparing the two species. The samples are two dissimilar with too many differentially expressed genes for gene ontology and pathway analysis.

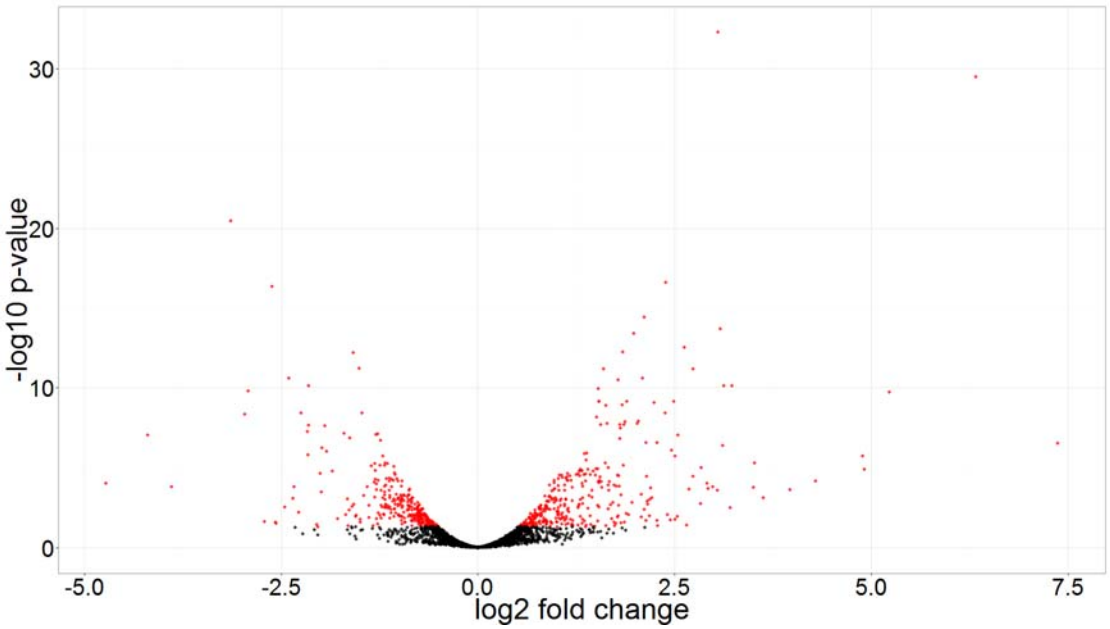


Figure 6.7 Large Number of Differentially Expressed Genes Between Species
A volcano plot showing the log2 fold change between the two species, Landrace and Jinhua, and the Benjamini-Hochberg adjusted p-value. Differentially expressed genes (p-value cutoff of 0.05) are shown in red.

6.3.3 Differentially Methylated Peaks

Differentially methylated peak regions were found using the windows from Figure 6.4 and the methods outlined previously in Chapter 4 Differentially Methylated Peak Regions (DMPRs). Using the *edgeR* Bioconductor package, TMM scaling was applied for the MeRIP-Seq and control samples separately,

and the subsequent adjusted log 2 enrichment distribution is shown in Figure 6.8, showing similar means of enrichment in the IPs across replicates. Volcano plots for differentially methylated peak regions in the soleus and tibialis anterior tissues are in Figure 6.9 and Figure 6.10, respectively. Multiple differentially methylated peak regions were found using a p-value cutoff of 0.10 and annotated to known genes.

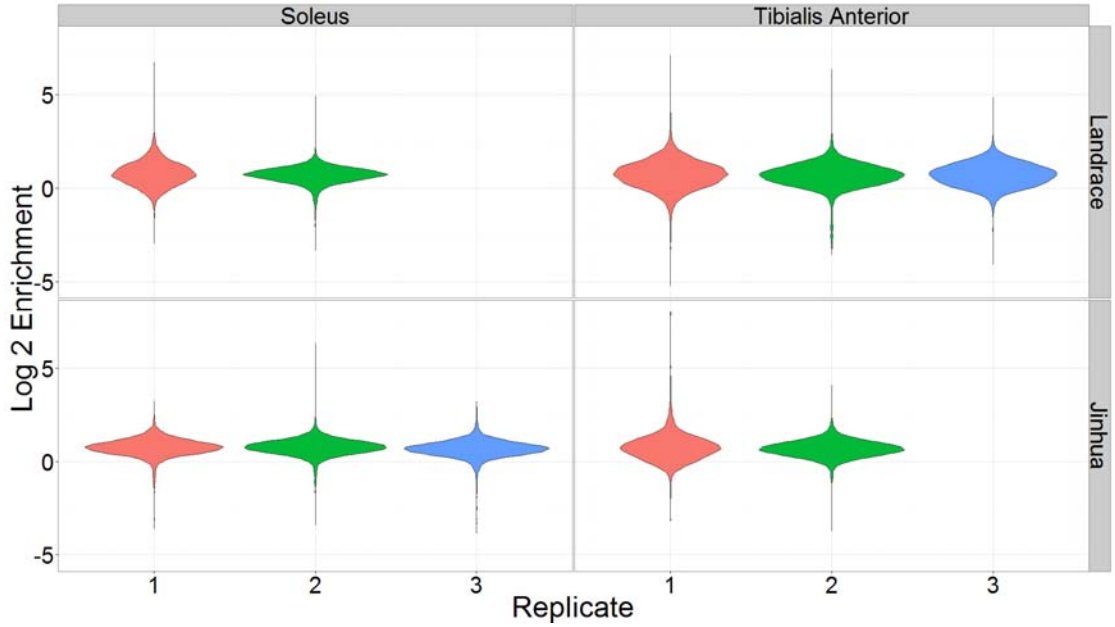


Figure 6.8 Adjusted Peak Enrichment Removes Technical Variance
Trimmed-Mean Method (TMM) Scaling shows the mean enrichment levels are normalized for differences observed earlier in IP efficiencies.

The soleus samples showed significant hypermethylation in gene annotations in the Jinhua relative to the Landrace, specifically in genes PCMT1, TPM2, RPL35, RPL27, RPL26, TNNC1, FDFT1, MYOZ1, RPL36, COX8H, TMOD4, BOLA1, RPL5, CALM3, GPANK1, RPS3A, and RPS4. The only gene that was hypomethylated was RN18S from ribosomal RNA. In contrast, there was significant hypomethylation in the tibialis anterior samples, especially in the

genes TPM2, RPL32, CAPZA2, RPL36, OAZ1, MDH2, TCTP, ITIH2, IDH2, and MB, while hypermethylation was observed in RN18S and RYR1. Detection of methylation changes in ribosomal RNAs are likely the result of ribosomal RNA contamination from the relatively poor polyA purification observed earlier.

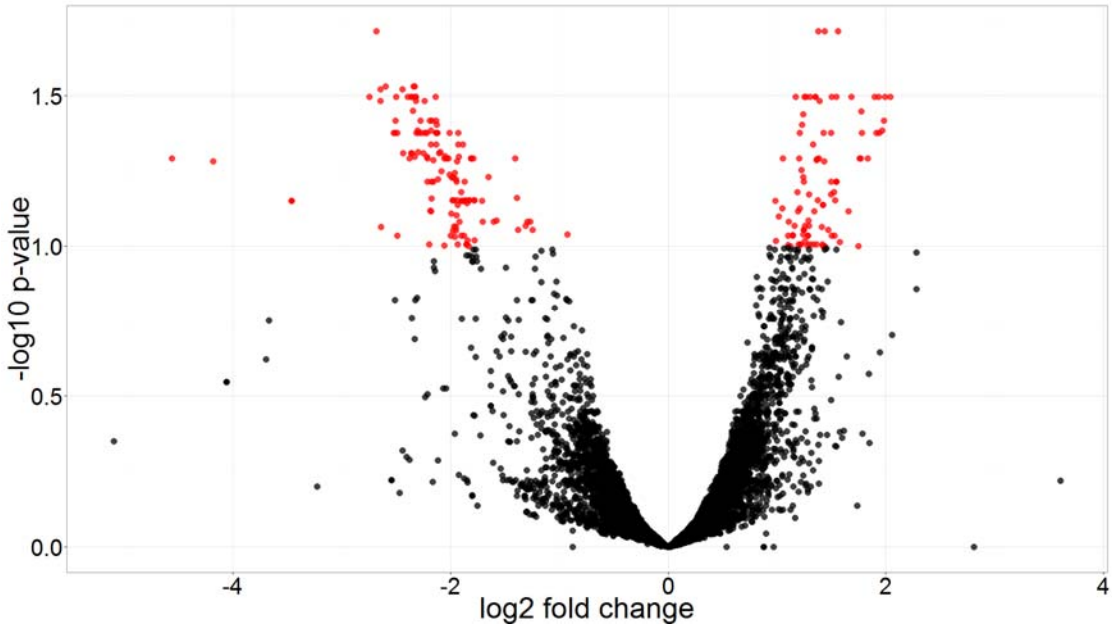


Figure 6.9 Differentially Methylated Peak Regions Between Species in Soleus Muscle
A volcano plot showing the log2 fold change and the Benjamini-Hochberg adjusted p-value between the Jinhua and Landrace species in the soleus muscle. Differentially expressed genes (p-value cutoff of 0.05) are shown in red.

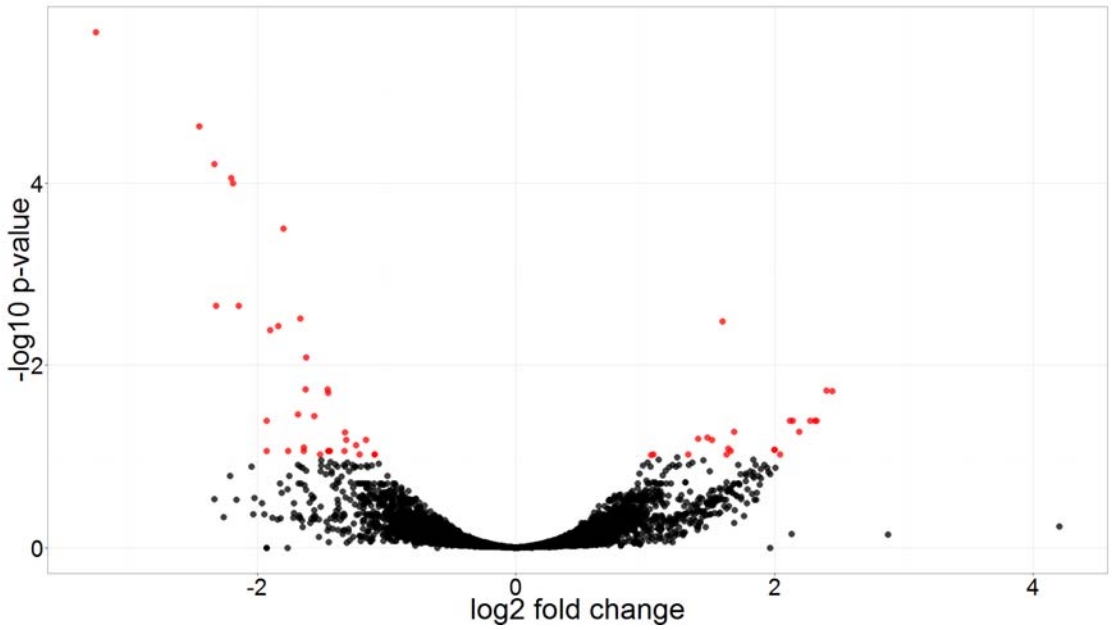


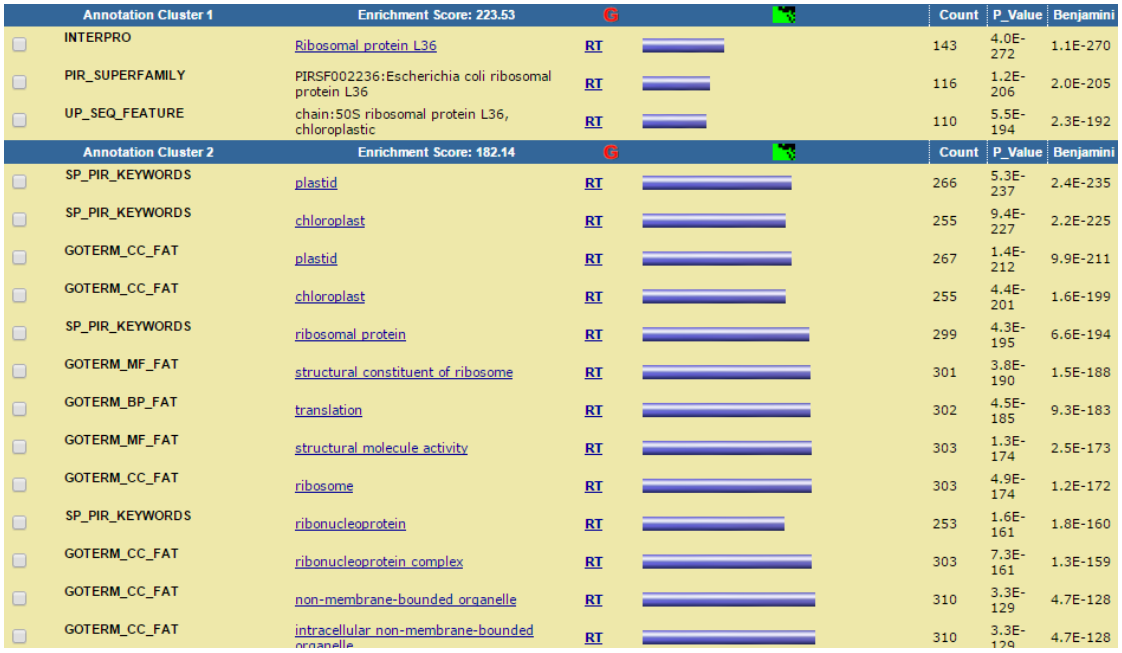
Figure 6.10 Differentially Methylated Peak Regions Between Species in Tibialis Anterior Muscle

A volcano plot showing the log2 fold change and the Benjamini-Hochberg adjusted p-value between the Jinhua and Landrace species in the tibialis anterior muscle. Differentially expressed genes (p-value cutoff of 0.05) are shown in red.

Annotation Cluster 1	Enrichment Score: ?			Count	P_Value	Benjamini
SP_PIR_KEYWORDS	ribosomal protein	RT		478	0.0E0	0.0E0
GOTERM_BP_FAT	translation	RT		449	4.9E-324	9.3E-322
GOTERM_MF_FAT	structural constituent of ribosome	RT		446	1.8E-294	8.7E-293
GOTERM_CC_FAT	ribosome	RT		485	1.1E-281	9.5E-280
GOTERM_MF_FAT	structural molecule activity	RT		448	2.3E-269	5.8E-268
SP_PIR_KEYWORDS	ribonucleoprotein	RT		391	2.5E-269	8.9E-268
GOTERM_CC_FAT	ribonucleoprotein complex	RT		485	6.7E-260	3.0E-258
SP_PIR_KEYWORDS	plastid	RT		318	6.1E-243	1.5E-241
GOTERM_CC_FAT	non-membrane-bounded organelle	RT		512	4.7E-235	1.4E-233
GOTERM_CC_FAT	intracellular non-membrane-bounded organelle	RT		512	4.7E-235	1.4E-233
SP_PIR_KEYWORDS	chloroplast	RT		292	8.2E-216	1.5E-214
GOTERM_CC_FAT	plastid	RT		318	1.7E-188	3.7E-187
GOTERM_CC_FAT	chloroplast	RT		292	1.7E-166	3.0E-165

Figure 6.11 Gene Ontology Analysis of Hypermethylated Genes in the Soleus

Gene ontology analysis using DAVID of hypermethylated genes from the soleus samples.



Annotation Cluster 1	Enrichment Score: 223.53			Count	P_Value	Benjamini
INTERPRO	Ribosomal protein L36	RT		143	4.0E-272	1.1E-270
PIR_SUPERFAMILY	PIRSF002236:Escherichia coli ribosomal protein L36	RT		116	1.2E-206	2.0E-205
UP_SEQ_FEATURE	chain:50S ribosomal protein L36, chloroplastic	RT		110	5.5E-194	2.3E-192
Annotation Cluster 2	**Enrichment Score: 182.14**			**Count**	**P_Value**	**Benjamini**
SP_PIR_KEYWORDS	plastid	RT		266	5.3E-237	2.4E-235
SP_PIR_KEYWORDS	chloroplast	RT		255	9.4E-227	2.2E-225
GOTERM_CC_FAT	plastid	RT		267	1.4E-212	9.9E-211
GOTERM_CC_FAT	chloroplast	RT		255	4.4E-201	1.6E-199
SP_PIR_KEYWORDS	ribosomal protein	RT		299	4.3E-195	6.6E-194
GOTERM_MF_FAT	structural constituent of ribosome	RT		301	3.8E-190	1.5E-188
GOTERM_BP_FAT	translation	RT		302	4.5E-185	9.3E-183
GOTERM_MF_FAT	structural molecule activity	RT		303	1.3E-174	2.5E-173
GOTERM_CC_FAT	ribosome	RT		303	4.9E-174	1.2E-172
SP_PIR_KEYWORDS	ribonucleoprotein	RT		253	1.6E-161	1.8E-160
GOTERM_CC_FAT	ribonucleoprotein complex	RT		303	7.3E-161	1.3E-159
GOTERM_CC_FAT	non-membrane-bounded organelle	RT		310	3.3E-129	4.7E-128
GOTERM_CC_FAT	intracellular non-membrane-bounded organelle	RT		310	3.3E-129	4.7E-128

Figure 6.12 Gene Ontology Analysis of Hypomethylated Genes in the Tibialis Anterior
Gene ontology analysis using DAVID of hypomethylated genes from the Tibialis Anterior samples.

6.4 Conclusion

Here we present the epitranscriptome of the porcine model, confirming the eukaryotic enrichment of m^6A sites near the stop codon. Global m^6A levels have been found to be anti-correlated with adipogenesis, in addition to m^6A sites required for adipogenesis. (Zhao et al., 2014) With further biological experiments comparing the knockdown of METTL3 or FTO, these DMPRs could have greater biological meaning and significance.

CHAPTER 7 CONCLUSION

7.1 Summary

Although RNA modifications have been known to exist for many years, only recently was the door opened to transcriptome-wide mapping methyl-6-adenosine. Shortly thereafter, the epitranscriptome expanded to include 5-methyl-cytidine in RNA, and the field continues to grow, exploring over 100 RNA modifications present. (Agris et al.; Cantara et al., 2011; Saletore et al., 2013) While we await the development of chemically-based methods, MeRIP-seq and MeRIPPeR combine an immunoprecipitation antibody-based enrichment protocol with a robust peak finder to find m^6A sites throughout the transcriptome. Each new epi-layer unveils a new level of transcriptional and translational regulation, and the full functional role of methyl-6-adenosine has yet to be discovered.

MeRIP-Seq and m^6A-seq are fully published protocols, with many new groups applying the same methods in new species and cell lines. In Chapter 2, I examined the protocol in more detail, identifying sources of bias and sample loss, such as the impact of ribosomal RNA contamination. The greatest challenge in m^6A experiments is the enormous input limit required, preventing analysis of clinical samples. Using a titration test, I determined not only the input limits of the protocol, but the consequences of using lower inputs, and how to achieve better IP enrichment.

In Chapter 3, I introduced MeRIPPeR as a MeRIP-Seq peak finder, designed to identify putative m^6A sites in the genome. Each step, from the decisions made during the MeRIP-Seq protocol to choices made on the computational side, has

an impact on the nature of the peaks called, and I explored the consequences of some of them, including the choice of aligner used, handling spliced data and spliced peaks, and correcting for ribosomal RNA contamination. I further explored biases that are introduced in the protocol, including the efficiency of the IP, and how that affects peak calling. I analyzed the data from the input titration test to compare the impact of using lower RNA inputs and the benefit of using two rounds of IP. Lastly, I compared the results and performance of MeRIPPeR against exomePeak, another m^6A-specific peak finder, and MACS2, and determined that MeRIPPeR called peaks the fastest and with the highest sensitivity.

I expanded on these methods and results in Chapter 4 to develop computational methods to identify differentially methylated peak regions (DMPRs). I discussed the impact of the efficiency of the IP, global changes in m^6A levels, and normalizing for changes in mRNA transcript levels. I developed a method building on existing methods to identify differentially expressed genes in RNA-sequencing data, using normalization factors to correct for technical variation in the IP. I compared the results with existing methods from exomePeak, though developing better methods of identifying DMPRs will require better validation.

In Chapter 5, I applied the methods I introduced in earlier chapters to try to understand the physiological and functional role of methyl-6-adenosine. Building on previous work with Ribavirin treatments affecting nuclear export of RNAs, as well as the well-characterized heat shock response, I implemented an experimental design to compare m^6A sites between total, nuclear, and cytosolic RNA. I first showed the results of the RNA-sequencing analysis, looking at differentially expressed genes and the up-regulation of the HSP70

gene family in response to heat shock. I then identified differentially methylated peak regions and found a correlation between methylation changes in intronic segments and the changes in the nuclear to cytosolic ratios of those genes, indicating that m^6A could play a role in nuclear export.

I then examined m^6A in the porcine model, mapping its epitranscriptome for the first time. Specifically looking at two breeds of pigs, the fattier Jinhua and the leaner Landrace, I examined how m^6A on the transcriptome level correlates with adipogenesis, identifying differentially expressed genes and differentially methylated genes and their associated gene ontology pathways.

7.2 Future Directions

Each month, new research unravels novel work in m^6A, a field that is gaining more attention. Science has followed Francis Crick's central dogma, focusing first on the genetic level, then building on the model to include epigenetic data, including histone modifications and DNA methylation, and now the world view is expanding to include the epitranscriptome. Whole 'omics studies that examine all of them simultaneously are rare, and expensive, but are the future of the field, working to build a more complete model of the complex biological regulatory networks. The difficulty with m^6A is that it is still a young field, most of which is still unknown. Each day we take small steps towards fully understanding its role, but without identifying all of the possible readers, writers, and erasers, we are still far from fully understanding the full functional relevance of m^6A.

Furthermore, DNA methylation patterns have been well-characterized in cancers, such as acute myeloid leukemia (AML), (Figueroa et al., 2010), with genetic mutations in IDH1 and IDH2 affecting DNA methylation patterns by disrupting the oxidative decarboxylation of isocitrate to alpha-keto-glutarate.

Tet2 catalyzes the conversion of 5-methylcytosine to 5-hydroxymethylcytosine, dependent on alpha-keto-glutarate. The RNA demethylases, FTO and ALKBH5, are also dependent on alpha-keto-glutarate for their function, and consequently, susceptible to IDH1 and IDH2 mutations in cancers. Examining this connection would demonstrate the clinical significance of the epitranscriptome, though research is hindered by the RNA input limit. The ability to barcode RNA fragments and pool samples together could open that door.

Lastly, in lieu of chemical-based methods to identify m^6A sites, third-generation single-molecule sequencers may be the future in identifying RNA modification. Instead of sequencing complementary DNA, introducing PCR and GC-biases, the native RNA strand could be sequenced, simultaneously reading in all modification data. I demonstrated a proof of principle on the Pac Bio RS, showing that direct RNA sequencing is not only possible, but has the ability to discern m^6A sites from adenosines. Oxford Nanopore Technologies' minION sequencers are smaller and easier to work with, and could be the future of all sequencing. Despite their high error rate, the protocol has the same potential to detect m^6A sites at single nucleotide resolution.

REFERENCES

Agris, P., Crain, P., Rozenski, J., Fabris, D., and Vendeix, F. The RNA Modification Database.

Akalin, A., Kormaksson, M., Li, S., Garrett-Bakelman, F.E., Figueroa, M.E., Melnick, A., and Mason, C.E. (2012). methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. Genome biology *13*, R87.

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome biology *11*, R106.

Assouline, S., Culjkovic-Kraljacic, B., Bergeron, J., Caplan, S., Cocolakis, E., Lambert, C., Lau, C.J., Zahreddine, H.A., Miller, W.H., Jr., and Borden, K.L. (2015). A phase I trial of ribavirin and low-dose cytarabine for the treatment of relapsed and refractory acute myeloid leukemia with elevated eIF4E. Haematologica *100*, e7-9.

Assouline, S., Culjkovic, B., Cocolakis, E., Rousseau, C., Beslu, N., Amri, A., Caplan, S., Leber, B., Roy, D.C., Miller, W.H., Jr.*, et al.* (2009). Molecular targeting of the oncogene eIF4E in acute myeloid leukemia (AML): a proof-of-principle clinical trial with ribavirin. Blood *114*, 257-260.

Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. Cell *129*, 823-837.

Bayley, H. (2006). Sequencing single molecules of DNA. Curr Opin Chem Biol *10*, 628-637.

Beemon, K., and Keith, J. (1977). Localization of N6-methyladenosine in the Rous sarcoma virus genome. J Mol Biol *113*, 165 - 179.

Benedict, C., Jacobsson, J.A., Ronnemaa, E., Sallman-Almen, M., Brooks, S., Schultes, B., Fredriksson, R., Lannfelt, L., Kilander, L., and Schioth, H.B. (2011). The fat mass and obesity gene is linked to reduced verbal fluency in overweight and obese elderly men. Neurobiology of aging *32*, 1159 e1151-1155.

Berulava, T., Rahmann, S., Rademacher, K., Klein-Hitpass, L., and Horsthemke, B. (2015). N6-Adenosine Methylation in MiRNAs. PloS one *10*, e0118438.

Bodi, Z., Button, J.D., Grierson, D., and Fray, R.G. (2010). Yeast targets for mRNA methylation. Nucleic acids research *38*, 5327-5335.

Bokar, J.A., Shambaugh, M.E., Polayes, D., Matera, A.G., and Rottman, F.M. (1997). Purification and cDNA cloning of the AdoMet-binding subunit of the human mRNA (N6-adenosine)-methyltransferase. Rna *3*, 1233-1247.

Borden, K.L., and Culjkovic-Kraljacic, B. (2010). Ribavirin as an anti-cancer therapy: acute myeloid leukemia and beyond? Leukemia & lymphoma *51*, 1805-1815.

Bringmann, P., and Luhrmann, R. (1987). Antibodies specific for N6-methyladenosine react with intact snRNPs U2 and U4/U6. FEBS letters *213*, 309 - 315.

Burkard, K.T., and Butler, J.S. (2000). A nuclear 3'-5' exonuclease involved in mRNA degradation interacts with Poly(A) polymerase and the hnRNA protein Npl3p. Molecular and cellular biology *20*, 604-616.

Cantara, W.A., Crain, P.F., Rozenski, J., McCloskey, J.A., Harris, K.A., Zhang, X., Vendeix, F.A., Fabris, D., and Agris, P.F. (2011). The RNA Modification Database, RNAMDB: 2011 update. Nucleic acids research *39*, D195-201.

Chen-Kiang, S., Nevins, J., and Darnell, J. (1979). N-6-methyl-adenosine in adenovirus type 2 nuclear RNA is conserved in the formation of messenger RNA. J Mol Biol *135*, 733 - 752.

Clancy, M., Shambaugh, M., Timpte, C., and Bokar, J. (2002). Induction of sporulation in Saccharomyces cerevisiae leads to the formation of N6-methyladenosine in mRNA: a potential mechanism for the activity of the IME4 gene. Nucleic acids research *30*, 4509 - 4518.

Crotty, S., Cameron, C., and Andino, R. (2002). Ribavirin's antiviral mechanism of action: lethal mutagenesis? Journal of molecular medicine *80*, 86-95.

Desrosiers, R., Friderici, K., and Rottman, F. (1974). Identification of methylated nucleosides in messenger RNA from Novikoff hepatoma cells. Proc Nat Acad Sci USA *71*, 3971 - 3975.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15-21.

Dominissini, D., Moshitch-Moshkovitz, S., Salmon-Divon, M., Amariglio, N., and Rechavi, G. (2013). Transcriptome-wide mapping of N(6)-methyladenosine by m(6)A-seq based on immunocapturing and massively parallel sequencing. Nature protocols *8*, 176-189.

Dominissini, D., Moshitch-Moshkovitz, S., Schwartz, S., Salmon-Divon, M., Ungar, L., Osenberg, S., Cesarkas, K., Jacob-Hirsch, J., Amariglio, N., Kupiec, M.*, et al.* (2012). Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. Nature *485*, 201-206.

Dubin, D.T., and Taylor, R.H. (1975). The methylation state of poly A-containing messenger RNA from cultured hamster cells. Nucleic acids research *2*, 1653-1668.

Eden, E., Lipson, D., Yogev, S., and Yakhini, Z. (2007). Discovering motifs in ranked lists of DNA sequences. PLoS computational biology *3*, e39.

Eden, E., Navon, R., Steinfeld, I., Lipson, D., and Yakhini, Z. (2009). GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC bioinformatics *10*, 48.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B.*, et al.* (2009). Real-time DNA sequencing from single polymerase molecules. Science *323*, 133 - 138.

Epstein, P., Reddy, R., Henning, D., and Busch, H. (1980). The nucleotide sequence of nuclear U6 (4.7 S) RNA. The Journal of biological chemistry *255*, 8901 - 8906.

Evdokimova, V., Ruzanov, P., Imataka, H., Raught, B., Svitkin, Y., Ovchinnikov, L.P., and Sonenberg, N. (2001). The major mRNA-associated protein YB-1 is a potent 5' cap-dependent mRNA stabilizer. The EMBO journal *20*, 5491-5502.

Fan, B., Du, Z.Q., and Rothschild, M.F. (2009). The fat mass and obesity-associated (FTO) gene is associated with intramuscular fat content and growth rate in the pig. Animal biotechnology *20*, 58-70.

Figueroa, M.E., Abdel-Wahab, O., Lu, C., Ward, P.S., Patel, J., Shih, A., Li, Y., Bhagwat, N., Vasanthakumar, A., Fernandez, H.F., *et al.* (2010). Leukemic IDH1 and IDH2 mutations result in a hypermethylation phenotype, disrupt TET2 function, and impair hematopoietic differentiation. Cancer cell *18*, 553-567.

Flusberg, B.A., Webster, D.R., Lee, J.H., Travers, K.J., Olivares, E.C., Clark, T.A., Korlach, J., and Turner, S.W. (2010). Direct detection of DNA methylation during single-molecule, real-time sequencing. Nature methods *7*, 461-465.

Fontanesi, L., Scotti, E., Buttazzoni, L., Dall'Olio, S., Bagnato, A., Lo Fiego, D.P., Davoli, R., and Russo, V. (2010). Confirmed association between a single nucleotide polymorphism in the FTO gene and obesity-related traits in heavy pigs. Molecular biology reports *37*, 461-466.

Fontanesi, L., Scotti, E., Buttazzoni, L., Davoli, R., and Russo, V. (2009). The porcine fat mass and obesity associated (FTO) gene is associated with fat deposition in Italian Duroc pigs. Animal genetics *40*, 90-93.

Frayling, T.M., Timpson, N.J., Weedon, M.N., Zeggini, E., Freathy, R.M., Lindgren, C.M., Perry, J.R., Elliott, K.S., Lango, H., Rayner, N.W., *et al.* (2007). A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. Science *316*, 889-894.

Fu, Y., Dominissini, D., Rechavi, G., and He, C. (2014). Gene expression regulation mediated through reversible m(6)A RNA methylation. Nature reviews Genetics *15*, 293-306.

Fustin, J.M., Doi, M., Yamaguchi, Y., Hida, H., Nishimura, S., Yoshida, M., Isagawa, T., Morioka, M.S., Kakeya, H., Manabe, I., *et al.* (2013). RNA-methylation-dependent RNA processing controls the speed of the circadian clock. Cell *155*, 793-806.

Gao, M., Fritz, D.T., Ford, L.P., and Wilusz, J. (2000). Interaction between a poly(A)-specific ribonuclease and the 5' cap influences mRNA deadenylation rates in vitro. Molecular cell *5*, 479-488.

Garrett-Bakelman, F.E., Sheridan, C.K., Kacmarczyk, T.J., Ishii, J., Betel, D., Alonso, A., Mason, C.E., Figueroa, M.E., and Melnick, A.M. (2015). Enhanced reduced representation bisulfite sequencing for assessment of DNA methylation at base pair resolution. Journal of visualized experiments : JoVE.

Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., *et al.* (2004). Bioconductor: open software development for computational biology and bioinformatics. Genome biology *5*, R80.

Gerken, T., Girard, C.A., Tung, Y.C., Webby, C.J., Saudek, V., Hewitson, K.S., Yeo, G.S., McDonough, M.A., Cunliffe, S., McNeill, L.A., *et al.* (2007). The obesity-associated FTO gene encodes a 2-oxoglutarate-dependent nucleic acid demethylase. Science *318*, 1469-1472.

Giannopoulou, E.G., and Elemento, O. (2011). An integrated ChIP-seq analysis platform with customizable workflows. BMC bioinformatics *12*, 277.

Gingras, A.C., Raught, B., and Sonenberg, N. (1999). eIF4 initiation factors: effectors of mRNA recruitment to ribosomes and regulators of translation. Annual review of biochemistry *68*, 913-963.

Guo, J., Shan, T., Wu, T., Zhu, L.N., Ren, Y., An, S., and Wang, Y. (2011). Comparisons of different muscle metabolic enzymes and muscle fiber types in Jinhua and Landrace pigs. Journal of animal science *89*, 185-191.

Harper, J., Miceli, S., Roberts, R., and Manley, J. (1990). Sequence specificity of the human mRNA N6-adenosine methylase in vitro. Nucleic acids research *18*, 5735 - 5741.

He, C. (2010). Grand challenge commentary: RNA epigenetics? Nature chemical biology *6*, 863-865.

Hess, M.E., Hess, S., Meyer, K.D., Verhagen, L.A., Koch, L., Bronneke, H.S., Dietrich, M.O., Jordan, S.D., Saletore, Y., Elemento, O., *et al.* (2013). The fat mass and obesity associated gene (Fto) regulates activity of the dopaminergic midbrain circuitry. Nature neuroscience *16*, 1042-1048.

Horowitz, S., Horowitz, A., and Nilsen, T. (1984). Mapping of N6-methyladenosine residues in bovine prolactin mRNA. Proc Nat Acad Sci USA *81*, 5667 - 5671.

Huang da, W., Sherman, B.T., and Lempicki, R.A. (2009a). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic acids research *37*, 1-13.

Huang da, W., Sherman, B.T., and Lempicki, R.A. (2009b). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature protocols *4*, 44-57.

Iwanami, Y., and Brown, G.M. (1968). Methylated bases of ribosomal ribonucleic acid from HeLa cells. Archives of biochemistry and biophysics *126*, 8-15.

Jaffrey, S.R. (2014). An expanding universe of mRNA modifications. Nature structural & molecular biology *21*, 945-946.

Jia, G., Fu, Y., Zhao, X., Dai, Q., Zheng, G., Yang, Y., Yi, C., Lindahl, T., Pan, T., Yang, Y.G., *et al.* (2011). N6-methyladenosine in nuclear RNA is a major substrate of the obesity-associated FTO. Nature chemical biology *7*, 885-887.

Johnson, D.S., Mortazavi, A., Myers, R.M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. Science *316*, 1497-1502.

Kane, S., and Beemon, K. (1985). Precise localization of m6A in Rous sarcoma virus RNA reveals clustering of methylation sites: implications for RNA processing. Molecular and cellular biology *5*, 2298 - 2306.

Kasianowicz, J.J., Brandin, E., Branton, D., and Deamer, D.W. (1996). Characterization of individual polynucleotide molecules using a membrane channel. Proceedings of the National Academy of Sciences of the United States of America *93*, 13770-13773.

Keller, L., Xu, W., Wang, H., Winblad, B., Fratiglioni, L., and Graff, C. (2011). The obesity related gene, FTO, interacts with APOE, and is associated with Alzheimer's disease risk: a prospective cohort study. J Alzheimers Dis *23*, 461 - 469.

Kentsis, A., Topisirovic, I., Culjkovic, B., Shao, L., and Borden, K.L. (2004). Ribavirin suppresses eIF4E-mediated oncogenic transformation by physical mimicry of the 7-methyl guanosine mRNA cap. Proceedings of the National Academy of Sciences of the United States of America *101*, 18105-18110.

Kentsis, A., Volpon, L., Topisirovic, I., Soll, C.E., Culjkovic, B., Shao, L., and Borden, K.L. (2005). Further evidence that ribavirin interacts with eIF4E. Rna *11*, 1762-1766.

Kong, H., Lin, L.F., Porter, N., Stickel, S., Byrd, D., Posfai, J., and Roberts, R.J. (2000). Functional analysis of putative restriction-modification system genes in the Helicobacter pylori J99 genome. Nucleic acids research *28*, 3216-3223.

Lahens, N.F., Kavakli, I.H., Zhang, R., Hayer, K., Black, M.B., Dueck, H., Pizarro, A., Kim, J., Irizarry, R., Thomas, R.S., *et al.* (2014). IVT-seq reveals extreme bias in RNA sequencing. Genome biology *15*, R86.

Lamond, A.I., and Spector, D.L. (2003). Nuclear speckles: a model for nuclear organelles. Nature reviews Molecular cell biology *4*, 605-612.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome biology *10*, R25.

Levis, R., and Penman, S. (1978). 5'-Terminal structures of poly(A)+ cytoplasmic messenger RNA and of poly(A)+ and poly(A) heterogeneous nuclear RNA of cells of the dipteran Drosophila melanogaster. J Mol Biol *120*, 487 - 515.

Lewis, J.D., and Izaurralde, E. (1997). The role of the cap structure in RNA processing and nuclear export. European journal of biochemistry / FEBS *247*, 461-469.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754-1760.

Li, S., Labaj, P.P., Zumbo, P., Sykacek, P., Shi, W., Shi, L., Phan, J., Wu, P.Y., Wang, M., Wang, C., *et al.* (2014a). Detecting and correcting systematic variation in large-scale RNA sequencing data. Nature biotechnology *32*, 888-895.

Li, S., and Mason, C.E. (2014). The pivotal regulatory landscape of RNA modifications. Annual review of genomics and human genetics *15*, 127-150.

Li, S., Tighe, S.W., Nicolet, C.M., Grove, D., Levy, S., Farmerie, W., Viale, A., Wright, C., Schweitzer, P.A., Gao, Y.*, et al.* (2014b). Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study. Nature biotechnology *32*, 915-925.

Li, Y., Song, S., Li, C., and Yu, J. (2013). MeRIP-PF: An Easy-to-use Pipeline for High-resolution Peak-finding in MeRIP-Seq Data. Genomics, Proteomics and Bioinformatics *11*, 72-75.

Lindquist, S., and Craig, E.A. (1988). The heat-shock proteins. Annual review of genetics *22*, 631-677.

Liu, J., Yue, Y., Han, D., Wang, X., Fu, Y., Zhang, L., Jia, G., Yu, M., Lu, Z., Deng, X.*, et al.* (2014). A METTL3-METTL14 complex mediates mammalian nuclear RNA N6-adenosine methylation. Nature chemical biology *10*, 93-95.

Meissner, A., Gnirke, A., Bell, G.W., Ramsahoye, B., Lander, E.S., and Jaenisch, R. (2005). Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. Nucleic acids research *33*, 5868-5877.

Meng, J., Cui, X., Rao, M.K., Chen, Y., and Huang, Y. (2013). Exome-based analysis for RNA epigenome sequencing data. Bioinformatics *29*, 1565-1567.

Meng, J., Lu, Z., Liu, H., Zhang, L., Zhang, S., Chen, Y., Rao, M.K., and Huang, Y. (2014). A protocol for RNA methylation differential analysis with MeRIP-Seq data and exomePeak R/Bioconductor package. Methods *69*, 274-281.

Meyer, K.D., Saletore, Y., Zumbo, P., Elemento, O., Mason, C.E., and Jaffrey, S.R. (2012). Comprehensive analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons. Cell *149*, 1635-1646.

Miao, Z.G., Wang, L.J., Xu, Z.R., Huang, J.F., and Wang, Y.R. (2009). Developmental changes of carcass composition, meat quality and organs in the Jinhua pig and Landrace. Animal : an international journal of animal bioscience *3*, 468-473.

Moss, B., Gershowitz, A., Stringer, J., Holland, L., and Wagner, E. (1977). 5' -Terminal and internal methylated nucleosides in herpes simplex virus type 1 mRNA. J Virol *23*, 234 - 239.

Munns, T.W., Liszewski, M.K., and Sims, H.F. (1977). Characterization of antibodies specific for N6-methyladenosine and for 7-methylguanosine. Biochemistry *16*, 2163-2168.

Nichols, J. (1979). N6-methyladenosine in maize poly(A)-containing RNA. Plant Sci Lett *15*, 357 - 361.

Nicol, J.W., Helt, G.A., Blanchard, S.G., Jr., Raja, A., and Loraine, A.E. (2009). The Integrated Genome Browser: free software for distribution and exploration of genome-scale datasets. Bioinformatics *25*, 2730-2731.

Perry, R.P., and Scherrer, K. (1975). The methylated constituents of globin mRNA. FEBS letters *57*, 73-78.

Ping, X.L., Sun, B.F., Wang, L., Xiao, W., Yang, X., Wang, W.J., Adhikari, S., Shi, Y., Lv, Y., Chen, Y.S., *et al.* (2014). Mammalian WTAP is a regulatory subunit of the RNA N6-methyladenosine methyltransferase. Cell research *24*, 177-189.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841-842.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic acids research.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139-140.

Robinson, M.D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. Genome biology *11*, R25.

Saletore, Y., Chen-Kiang, S., and Mason, C.E. (2013). Novel RNA regulatory mechanisms revealed in the epitranscriptome. RNA biology *10*, 342-346.

Saletore, Y., Meyer, K., Korlach, J., Vilfan, I.D., Jaffrey, S., and Mason, C.E. (2012). The birth of the Epitranscriptome: deciphering the function of RNA modifications. Genome biology *13*, 175.

Schwartz, S., Bernstein, D.A., Mumbach, M.R., Jovanovic, M., Herbst, R.H., Leon-Ricardo, B.X., Engreitz, J.M., Guttman, M., Satija, R., Lander, E.S., *et al.* (2014). Transcriptome-wide mapping reveals widespread dynamic-regulated pseudouridylation of ncRNA and mRNA. Cell *159*, 148-162.

SEQC MACQ-III Consortium (2014). A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. Nature biotechnology *32*, 903-914.

Seumois, G., Chavez, L., Gerasimova, A., Lienhard, M., Omran, N., Kalinke, L., Vedanayagam, M., Ganesan, A.P., Chawla, A., Djukanovic, R., *et al.* (2014). Epigenomic analysis of primary human T cells reveals enhancers associated with TH2 memory cell differentiation and asthma susceptibility. Nature immunology *15*, 777-788.

Shishkin, A.A., Giannoukos, G., Kucukural, A., Ciulla, D., Busby, M., Surka, C., Chen, J., Bhattacharyya, R.P., Rudy, R.F., Patel, M.M., *et al.* (2015). Simultaneous generation of many RNA-seq libraries in a single reaction. Nature methods *12*, 323-325.

Song, C., Clark, T., Lu, X., Kislyuk, A., Dai, Q., Turner, S., He, C., and Korlach, J. (2011). Sensitive and specific single-molecule sequencing of 5-hydroxymethylcytosine. Nature methods *9*, 75 - 77.

Song, C.X., Yi, C., and He, C. (2012). Mapping recently identified nucleotide variants in the genome and transcriptome. Nature biotechnology *30*, 1107-1116.

Squires, J.E., Patel, H.R., Nousch, M., Sibbritt, T., Humphreys, D.T., Parker, B.J., Suter, C.M., and Preiss, T. (2012). Widespread occurrence of 5-methylcytosine in human coding and non-coding RNA. Nucleic acids research *40*, 5023-5033.

Trapnell, C., Pachter, L., and Salzberg, S. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105 - 1111.

Veliz, E., Easterwood, L., and Beal, P. (2003). Substrate analogues for an RNA-editing adenosine deaminase: mechanistic investigation and inhibitor design. J Am Chem Soc *125*, 10867 - 10876.

Visa, N., Izaurralde, E., Ferreira, J., Daneholt, B., and Mattaj, I.W. (1996). A nuclear cap-binding complex binds Balbiani ring pre-mRNA cotranscriptionally and accompanies the ribonucleoprotein particle during nuclear export. The Journal of cell biology *133*, 5-14.

Wang, X., Lu, Z., Gomez, A., Hon, G.C., Yue, Y., Han, D., Fu, Y., Parisien, M., Dai, Q., Jia, G., *et al.* (2014a). N6-methyladenosine-dependent regulation of messenger RNA stability. Nature *505*, 117-120.

Wang, Y., Li, Y., Toth, J.I., Petroski, M.D., Zhang, Z., and Zhao, J.C. (2014b). N6-methyladenosine modification destabilizes developmental regulators in embryonic stem cells. Nature cell biology *16*, 191-198.

Wei, C.-M., Gershowitz, A., and Moss, B. (1975). Methylated nucleotides block 5' terminus of HeLa cell messenger RNA. Cell *4*, 379 - 386.

Wei, C.-M., and Moss, B. (1977a). Nucleotide sequences at the N6-methyladenosine sites of HeLa cell messenger ribonucleic acid. Biochemistry *16*, 1672 - 1676.

Wei, C., Gershowitz, A., and Moss, B. (1976). 5'-Terminal and internal methylated nucleotide sequences in HeLa cell mRNA. Biochemistry *15*, 397 - 401.

Wei, C.M., and Moss, B. (1977b). Nucleotide sequences at the N6-methyladenosine sites of HeLa cell messenger ribonucleic acid. Biochemistry *16*, 1672-1676.

Will, C.L., and Luhrmann, R. (2011). Spliceosome structure and function. Cold Spring Harbor perspectives in biology *3*.

Wu, T.D., and Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics *26*, 873-881.

Yamamoto, S., Wu, Z., Russnes, H.G., Takagi, S., Peluffo, G., Vaske, C., Zhao, X., Moen Vollan, H.K., Maruyama, R., Ekram, M.B., *et al.* (2014). JARID1B is a luminal lineage-driving oncogene in breast cancer. Cancer cell *25*, 762-777.

Yang, J., Loos, R.J., Powell, J.E., Medland, S.E., Speliotes, E.K., Chasman, D.I., Rose, L.M., Thorleifsson, G., Steinthorsdottir, V., Magi, R., *et al.* (2012). FTO genotype is associated with phenotypic variability of body mass index. Nature *490*, 267-272.

Zhang, L., Meng, J., Liu, H., Cui, X., Zhang, S.-W., Chen, Y., and Huang, Y. (2014a). Detecting differentially methylated mRNA from MeRIP-Seq with likelihood ratio test.

Paper presented at: 2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP).

Zhang, Y.-C., Zhang, S.-W., Liu, L., Zhang, L., Liu, H., Cui, X., Huang, Y., and Meng, J. (2014b). Differential analysis of RNA methylome with improved spatial resolution. Paper presented at: 2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP).

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., *et al.* (2008). Model-based analysis of ChIP-Seq (MACS). Genome biology *9*, R137.

Zhao, X., Yang, Y., Sun, B.F., Shi, Y., Yang, X., Xiao, W., Hao, Y.J., Ping, X.L., Chen, Y.S., Wang, W.J., *et al.* (2014). FTO-dependent demethylation of N6-methyladenosine regulates mRNA splicing and is required for adipogenesis. Cell research *24*, 1403-1419.

Zheng, G., Dahl, J.A., Niu, Y., Fedorcsak, P., Huang, C.M., Li, C.J., Vagbo, C.B., Shi, Y., Wang, W.L., Song, S.H., *et al.* (2013). ALKBH5 is a mammalian RNA demethylase that impacts RNA metabolism and mouse fertility. Molecular cell *49*, 18-29.